Remarks

Claims 12, 18-20, 23 and 26-27 have been amended; claims 37-43 have been added. Claims 1-11, 17, 25 and 28-32 are cancelled. Upon entry of this amendment, claims 12-16, 18-24, 26-27, and 33-43 will be pending.

Support for the amendments of claims 12 and 20 can be found, *e.g.*, in paragraph 8 at page 3; paragraph 11 at page 4; paragraph 16 at page 5; paragraph 28 at page 8, of the specification. Claim 20 has been further amended to incorporate the limitations of previously presented claims 23-24. Support for the amendments to claim 19 and 27 can be found, *e.g.*, in paragraph 17, page 5 of the specification. Support for the amendments to claims 15 and 23 can be found, *e.g.*, in paragraph 33. Support for the new claims can be found, *e.g.* in paragraph 16, page 5; in paragraph 46 bridging pages 15-16; and in paragraph 8 at page 6, of the specification. No new matter has been added.

Applicants hereby submit a revised Sequence Listing in paper and computer readable form, together with a verification statement to fulfill the requirements of 37 C.F.R. §1.821-1.825. The amendments in the specification replace the previous Sequence Listing with the paper copy of the Sequence Listing submitted herein. The specification is further amended to insert new sequence identifiers. No new matter has been added to the specification by this submission.

The claim amendments and cancellations made herein have been made solely to expedite prosecution of the instant application and should not be construed as an acquiescence to any of the Examiner's rejections.

Sequence Compliance

In paragraph 5 of the Office Action, the Office has objected to the sequence disclosed in Figure 1A-1EEE of the specification as requiring an appropriate sequence identifier and inclusion in the Sequence Listing.

In response, the specification has been amended to properly reference the sequence disclosed in Figure 1A-1EEE. In addition, a new sequence listing disclosing the amino acid sequence in the Figure and supporting statements in compliance with 37 C.F.R. 1.821 are being submitted herewith. Reconsideration and withdrawal of this objection is respectfully requested.

Rejection of claims 12-16, 18-24, 26-27, and 33-34 for lack of written description under 35 U.S.C. §112, first paragraph

In paragraph 6 (pages 3-8) of the Office Action, the Office has maintained the rejection of claims 12-16, 18-24, 26-27, and 33-34 under 35 U.S.C. §112, first paragraph for allegedly lack of written description under 35 USC §112, first paragraph. Each aspect asserted in support of maintaining this rejection is addressed individually below.

In one aspect, the Office maintains the rejection of the pending claims for lack of written description because "the 3-D models of BACE encompass widely variant 3-D models of BACE having any sequence of amino acids and any tertiary structure as long as the resulting structure has the recited active site amino acids." To support this conclusion, the Office states that:

> [T]he specification acknowledges that the structural coordinates can be "modified from those of Figures 1A-1EEE by mathematical manipulation." Such "mathematical manipulation" has been interpreted as encompassing mathematical manipulation of the data of Figures 1A-1EEE by an algorithm to generate a homology model. ....Thus, contrary to applicant's position, the 3-D models of BACE encompass widely variant 3-D models of BACE having any sequence of amino acids and any tertiary structure as long as the resulting structure has the recited active site amino acids. In this case, it is highly unpredictable as to whether the resulting homology model(s) will maintain a conformation of a catalytically active BACE polypeptide, which is evidence by Flower ("Drug Design, Cutting Edge Approaches," Royal Society of Chemistry, Cambridge, UK, 2002), which, addressing the use of homology models for identifying lead drugs, discloses "[p]roblems still exist, however: the fitting together of protein domains in a multi-domain protein, the determination of the most likely conformation of protein loops, the correct positioning of amino acid side chains, flexible ligand docking – to name only a few" (p. 25, middle). (page 4 of Office Action)

To expedite prosecution of the instant application, independent claims 12 and 20 have been amended, *inter alia*, to specify that the BACE sequence used in the claimed methods includes, or consists essentially of, the amino acid sequence of residues 58-477 of human BACE (SEQ ID NO:1), and to recite the particular structural coordinates of numerous (or all) residues of the human BACE sequence according to Figures 1A-1EEE ± the root mean square deviation specified. Claim 12, as amended, and newly added claims 38-39, which depend from claim 20, specify that the BACE three dimensional structure further comprises the structural coordinates of an APP peptide inhibitor (*e.g.*, an APP peptide inhibitor having the sequence specified as recited by claims 12 and 39 as amended herein). Newly added claims 42-43 further include the step of providing a crystalline composition of BACE having a particular group symmetry and unit parameter. Therefore, the pending claims, as amended herein, are directed to methods of identifying agents that interact with BACE by providing a three dimensional structure of BACE based on the particular structural coordinates of the human BACE sequence specified (e.g., the structural coordinates of human $BACE_{58-447}$ complexed with a particular APP peptide); generating a three dimensional model using the aforesaid three dimensional structure; performing computer fitting analysis to determine the degree of association between BACE and a candidate agent; and identifying the agent. The claims, as amended herein, do not encompass the "widely variant 3-D models of BACE having any sequence of amino acids and any tertiary structure," as alleged by the Office.

Applicants respectfully traverse the Office's position stating that the mathematical manipulations of the structural coordinates of the selected amino acids in Figures 1A-1EEE specified in the claims encompass "highly unpredictable" homology modeling. From the outset, the term "relative structural coordinates," as used in the claims refers to a term-of-art describing the relationship between the atoms in the three dimensional structure. It is clarified for the record that the relationship between the atoms in the 3-D structure remains the same, but the entire three dimensional configuration can be shifted, *e.g.*, by rotation or inversion, integer addition or subtraction. Thus, although the actual number of the coordinates may shift, the relationship between the atoms stays the same. This is explained in the specification in paragraph 21 as follows:

> The structural coordinates of the present invention may be modified from the original set provided in Figure 1 by mathematical manipulation, such as by inversion or integer addition or subtractions. As such, it is recognized that the structural coordinates of the present invention are relative, and are in no way specifically limited by the actual x, y, z coordinates of Figure 1. (Specification page 6, paragraph 21).

The degree of deviation from the three dimensional coordinates used in the claimed methods is provided by the "root mean square deviation" defined in the instant specification in paragraph 23. The pending claims require a limited deviation from the backbone atoms of BACE up to 1.5Å. Thus, the coordinates encompassed by the three dimensional models in Figures 1A-1EEE, as claimed, cannot be any structural coordinates, but only those within the root mean square deviation parameters specified.

With respect to the Office's statement regarding the high level of unpredictability of homology modeling and the reference by Flower ("Drug Design, Cutting Edge Approaches," Royal Society of Chemistry, Cambridge, UK, 2002) allegedly cited in support of this statement, Applicants clarify that the claims do not encompass homology models that create *de novo* structures based on structural data available on a homologous sequence (*e.g.*, a homologous aspartic protease), in the absence of the solved BACE crystal structure. In the present application, the three-dimensional structure used to generate the BACE three dimensional models encompassed by the claims is disclosed in Figures 1A-1EEE. The claims are directed to screening methods that are based on the structural coordinates of the particular human BACE sequences within the particular deviation parameters from the backbone atoms of BACE specified by the claims. ***Thus, homology modeling, without the benefit of the solved structure as described in Flower, is not relevant to the present claims.***

Even if homology models as described by the Flower reference were relevant to the claimed invention (which Applicants submit that is not relevant), the teachings of the Flower reference are clarified below for the record. The reference by Flower describes two elements of structure-based design, *i.e.*, the traditional model which involves the solution of the protein structure followed by computational strategies to predict or identify putative ligands, and a second model based on homology modeling. As to homology modeling, the reference states that:

> The other part of structure-based design is homology modeling. Here the structure of a protein is modeled using the experimentally determined structure of homologous programs. It is now a well-established technique and automated methods that remove much of the tedium from the routine production of such models are now well known [citing Composer and Modeller, two commercially available computer models for modeling]. Problems still exist, however: the fitting together of protein domains in a

> multi-domain protein, the determination of the most likely conformation of protein loops, the correct positioning of amino acid side chains, flexible ligand docking – to name only a few [quoted above] (page 25 of the Flower reference (emphasis added)).

Thus, the Flower reference describes two alternative approaches for generating three dimensional models – a traditional approach that involves generating the protein crystal and X-ray diffraction, which is exemplified in the instant application (*see e.g.*, paragraphs 41-46), and a second approach involving homology modeling. Both approaches were established techniques at the time of filing, as acknowledged by the reference: "[Homology modeling] is now a well-established technique and automated methods that remove much of the tedium from the routine production of such models are now well known." (Flower, *supra* at page 25).

In view of the amendment to the pending claims to recite the particular BACE amino acid sequence and structural coordinates specified, Applicants respectfully submit that the claims, as presently amended, provide sufficient structural and functional characteristics in common of the BACE three dimensional model used in the claimed methods to show that Applicants were in possession of the claimed genus at the time of filing, and thus fully comply with the written description requirement.

In another aspect of this rejection (page 5 of the Office Action), the Office rejects claims 16, 19, 24 and 27 in their recitation of the term "BACE" and "APP," by stating that "the specification indicates structural preferences for the genus of BACE polypeptides" and that these "'preferences" are non-limiting and thus the genus of BACE polypeptides encompasses any polypeptide having the ability to cleave APP at residue 671."

This aspect of the rejection has been met by amending the pending claims to specify that the BACE used in the claimed methods include, or consist essentially of, the amino acid sequence of residues 58-447 of human BACE (SEQ ID NO:1). Thus, as amended, the claims do not encompass "any polypeptide having the ability to cleave APP at residue 671," but instead provide sufficient structural and functional characteristics commensurate with the scope of the claims. Moreover, the sequence and functional characterization of BACE were well known in the art at the time the application was filed (*see e.g.*, Sinha, S. *et al.* (1999) *Nature* 402:537-540; Yan, R. *et al.* (1999) *Nature* 402:533-537; Hussain, I. *et al.* (1999) *Mol. Cell. Neurosci.* 14(6):419-427). Given the knowledge in the art and the teachings of the specification, a skilled practitioner at the filing date would have recognized that Applicants were in possession of, and had adequately described, the BACE genus recited in the pending claims.

As to the term "APP," the Office contends that the specification acknowledged that the genus encompass conservative mutants of APP having accession number CAA31380, and that the revision history for accession number CAA31380 has been revised numerous times and it is unclear as to which sequence is referred to by "accession number CAA31380."

Applicants traverse this aspect of the rejection. Amyloid beta precursor proteins, APP^{770} and its isoforms APP^{695} and APP^{751} were well known and characterized in the art at the time the application was filed (*see e.g.*, Kang, J. *et al.* (1987) *Nature* 325:7330736 and Lemaire, H-G *et al.* (1989) *Nucleic Acid Res.* 17:517-522 (describing isoform APP^{695}); Ponte, P. *et al.* (1988) *Nature* 331:525-527 (describing isoform APP^{751}); Saito, M. *et al.* (1997) *Nucleic Acid Res.* 25(9):1802-1808 (describing isoform APP^{770}) (copies submitted herewith as Exhibits A-D, respectively). The specification discloses the biology and chacterization of the APP 770 amino acid isoform, *see e.g.*, paragraph 3 at page 1; paragraph 17 at page 5, and paragraphs 55-58 at pages 18 through 20, of the specification. The sequence of an APP peptide (SEVNStaVAEF, wherein Sta is (S)-Statine) and its relationship to the APP751 Swedish family mutation is described in paragraphs 18, 46, 55 and 58 of the specification.

The Office has taken issue with the incorporation in the specification of CAA31830 as describing one of the isoforms of APP. CAA31830 provides the amino acid sequence of human APP^{695}. The submission of June 23, 1995 (submitted herewith as Exhibit E) indicates the following: (1) the version of the sequence provided on June 23, 1995 replaced sequence version gi:35600; and (2) provides a cross-reference to a contemporaneous paper by the same author, Lemaire, H-G *et al.* (1989) *Nucleic Acid Res.* 17:517-522 (supra). The revision history of CAA31830 indicates that the sequence has not been modified since June 23, 1995. In this regard, Applicants submit a comparison of the CAA31830 submissions from June 23, 1995 and November 14, 2006 (most recent revision of this entry), which notes the changes added to the November 14 submission compared to the June 23 submission in blue, while the identical text is noted in black (see Exhibit F, submitted herewith). As shown in Exhibit F, information concerning the reference to Lemaire, H-G *et al.* (1989) *Nucleic Acid Res.* 17:517-522 (supra) and the APP^{695} amino acid sequence disclosed was identical in the submission of June 23, 1995 and

November 14, 2006. Thus, there is no indication in the revision history of a revision of the APP^{695} sequence available at the time of filing of the present application. The availability of the sequence in an NCBI database, as opposed to a published paper in a paper volume of a journal, in no way detracts from the fact that one of ordinary skill in the art would have had, and currently has, access to the APP^{695} disclosed at the time of filing of the instant application, and the amino acid sequence disclosed in previous and current submissions has not changed since June 23, 1995.

Thus, based on the knowledge in the art regarding APP and the descriptions provided in the specification, a skilled practitioner at the filing date would have recognized that Applicants were in possession of, and had adequately described, the APP genus recited in the pending claims. As the Federal Circuit clearly articulated in *Capon v. Eshhar* (418 F.3d 1349) "a requirement that patentees recite known DNA structures, if one existed, would serve no goal of the written description requirement." *Id.* More specifically, the Federal Circuit, citing the *Capon* decision, stated that:

> As we stated in *Capon*, "[t]he 'written description' requirement states that the patentee must describe the invention; it does not state that every invention must be described in the same way. As each field evolves, the balance also evolves between what is known and what is added by each inventive contribution." (citation omitted). Indeed, the forced recitation of known sequences in patent disclosures would only add unnecessary bulk to the specification. Accordingly we hold that where, as in this case, accessible literature sources clearly provided, as of the relevant date, genes and their nucleotide sequences (here "essential genes"), satisfaction of the written description requirement does not require either the recitation or incorporation by reference...of such genes and sequences. *Id.*

Lastly, the Office states on page 6-8 of the Office Action that:

> [t]he specification discloses only a single representative species of the genus of claimed crystals, *i.e.*, a crystal of the BACE protein prepared as disclosed at pp. 14-1 5 of the specification in complex with inhibitor SVENStaVAEF having the space group symmetry 1222 and having vector lengths a=86.627, b=130.861 Å, and c=130.729 Å and a=β= y=90° and the specification discloses only a single representative species of the genus of crystallization methods, *i.e.*, the method disclosed at pp. 14-1 5, paragraphs 41-44 of the specification. Other than these single disclosed species, the specification fails to describe any additional representative species of the recited genus, which encompasses widely variant species, including crystals of polypeptides that are widely variant in

polypeptide sequence, space group, and unit cell dimensions that are unliganded or have any bound ligand, are produced by essentially any method of crystallization.

Applicants respectfully submit that this aspect of the rejection has been met in view of the claim amendments discussed above, the substance of which is reiterated here. The claims, as amended herein, are directed to method of using three dimensional models based on the particular structural coordinates and human BACE sequences specified. Since the structural coordinates are based on X-ray diffraction of the particular BACE crystals disclosed, the structural coordinates specified by the claims are a direct result from the particular BACE crystals disclosed. Newly added claims 42-43 add the further step of "providing a crystalline composition of BACE, wherein the crystalline composition had space group I222, and unit cell parameters a=86.627, b=130.861, c=130.729, and α=β=γ=90°." Applicants submit that the present rejection would not apply to the newly added claims.

The sequence and functional characterization of BACE was described in the specification and was well known in the art at the time the application was filed (*see e.g.*, Exhibit A, discussed above). The recitation of the space group and unit cell parameters specified by claims 42-43 provide the requisite structural features in common to satisfy the written description requirement, as analogized from the hypothetical crystal claim 1 exemplified in case 4 of the "Trilateral Project WM4 Comparative Studies in New Technologies: Report on Comparative Study on Protein 3-Dimensional (3-D) Structure Related Claims" released in November 2002 ("the Trilateral Report"). The USPTO indicated in the Trilateral Report that hypothetical claim 1 would meet the written description requirement because the crystal structure of the protein is provided in the claim by specifying the cell unit dimension. More specifically, hypothetical claim 1 in case 4 of the Trilateral Report is directed to a crystalline form of a known protein P, and reads as follows: "A crystalline form of protein P having unit cell dimensions of a=4.0nm, b=7.8nm, and c=11.0nm." At pages 8 and 66 of the report, the hypothetical specification of case 4 is described as including, *inter alia*, that the inventors have newly produced a stable crystalline form of protein P and that the description gives experimental data with explanations of how to make the crystals. The Trilateral Report, at page 67, and referring to the claim of case 4, states that "the claim complies with the written description requirement because the structure of protein

P is provided." Thus, reconsideration and withdrawal of this aspect of the rejection is respectfully requested.

Given the knowledge in the art and the teachings of the specification, a skilled practitioner at the filing date would have recognized that Applicants were in possession of, and had adequately described, the genus recited in the pending claims to satisfy the standard set forth in the MPEP (*see, e.g.*, MPEP § 2163(II)(A)(3)(ii)). Accordingly, Applicants submit that the claims, as presently pending, fully satisfy the written description requirement. Reconsideration of this rejection is respectfully requested.

Rejection of claims 12-16, 18-24, 26-27, and 33-34 for lack of enablement under 35 U.S.C. 112, first paragraph

In paragraph 7 of the Office Action (pages 8-17), the Office has rejected the pending claims for alleged lack of enablement. Since a number of the grounds for this rejection parallel the written description rejection discussed above, Applicants reiterate their position above and will focus the response as it applies to the present enablement rejection.

In one aspect of this rejection, the Office restates the position that "the claims broadly encompass the use of 3-D models of BACE having any sequence of amino acids and any tertiary structure as long as the resulting structure has the recited active site amino acids."

As stated above, this aspect of the rejection is now moot in view of the claim amendments to specify that the BACE sequence used in the claimed methods includes, or consists essentially of, the amino acid sequence of residues 58-477 of human BACE (SEQ ID NO:1), and to recite the particular structural coordinates of numerous (or all) residues of the human BACE sequence according to Figures 1A-1EEE ± the root mean square deviation specified. Claim 12, as amended, and newly added claims 38-39, which depend from claim 20, specify that the BACE three dimensional structure further comprises the structural coordinates of an APP peptide inhibitor (*e.g.*, an APP peptide inhibitor having the sequence specified as recited by claims 12 and 39 as amended herein). Newly added claims 42-43 further include the step of providing a crystalline composition of BACE having a particular group symmetry and unit parameter. Therefore, the pending claims, as amended herein, are directed to methods of identifying agents that interact with BACE by providing a three dimensional structure of BACE based on the particular structural coordinates of the human BACE sequence specified; generating a three dimensional model using the aforesaid three dimensional structure; performing computer fitting analysis to determine the degree of association between BACE and a candidate agent; and identifying the agent. The claims, as amended herein, do not encompass the "widely variant 3-D models of BACE having any sequence of amino acids and any tertiary structure," as alleged by the Office.

In response to the Office's comments regarding the use of the term "relative structural coordinates" in the claims, Applicants reiterate the discussion above regarding this term. As used in the specification and understood in the art, the term "relative structural coordinates" requires the same relationship between the atoms in the 3-D structure, but allows for a change in the 3-D configuration, *e.g.*, by rotation or inversion, integer addition or subtraction. Thus, although the actual number of the coordinates may shift, the relationship between the atoms stays the same.

As to the breadth of the claims, the method claims, as presently amended, require the use of human BACE having the amino acid sequence specified. The degree of deviation from the three dimensional coordinates used in the claimed methods is limited to a deviation from the backbone atoms of human BACE sequence specified up to 1.5Å. Thus, the coordinates encompassed by the three dimensional models in Figures 1A-1EEE as claimed cannot be any structural coordinates, but only those having a particular sequence within the root mean square deviation parameters specified.

Moreover, the amino acid sequence and domain characterization of BACE were known in the art at the time the instant application was filed and are extensively described in the instant application. For example, a detailed characterization of the location of, and interactions between, residues and domains of BACE, and how these correlate with biological activity is provided, *e.g.*, starting in paragraph 47 through 58 of the instant application. The location of the active site of BACE was also disclosed in the application. Selected residues important for substrate interaction are disclosed and claimed in the present application. Therefore, at the time the present application was filed, one of ordinary skill in the art would have known to make changes preferentially outside of these selected amino acid stretches of human BACE. Techniques for generating mutant BACE proteins were well known in the art at the time the present application was filed and were performed routinely by molecular biologists. Similarly, crystallization and molecular modeling techniques are extensively described in the instant application (see, *e.g.*, in the Example starting at page 15, paragraph 46), and were known in the art at the time the application was filed. Software systems for generating three-dimensional models were also described in the specification (*see e.g.*, page 10, paragraph 33), and were known in the art. Therefore, Applicants submit that following the teachings of the specification, one of ordinary skill in the art would have been able to generate BACE 3-D models having the sequence specified by practicing routine experimentation.

The Office reiterates its position on pages 10-11 of the Office Action that the mathematical manipulations of the structural coordinates of the selected amino acids in Figure 1 specified in the claims encompass "highly unpredictable" homology modeling, and cites to the Flower reference (discussed above) and Lambert *et al.* US 04/0137518, as supporting the notion that current homology models cannot provide the necessary degree of specificity to design modulators.

In response, Applicants reiterate the discussion above regarding the Flower reference. Methods of generating three dimensional models by experimental or homology models were known techniques at the time the present application was filed, and not a "highly unpredictable" art as characterized by the Office. It is important to emphasize that ***once the structural coordinates of a protein, in this case human BACE, are identified*** as set forth in Figures 1A-1EEE of the application, three-dimensional models within the requisite deviation from the backbone atoms of BACE could have been generated by one of ordinary skilled in the art without undue experimentation. Software systems for generating three-dimensional models were also described in the specification, and were known in the art at the time the application was filed. This is completely in agreement with the Lambert reference cited by the Examiner, which provides in paragraph 17 (relied upon by the Office on page 9 of the Office Action to support this rejection) that:

> [0017] The solved PPAR.alpha.-ligand crystal structure would provide structural details and insights necessary to design a modulator of PPAR.alpha. that maximizes preferred requirements for any modulator, i.e. potency and specificity. By exploiting the structural details obtained from a PPAR-ligand crystal structure, it would be possible to design a PPAR modulator that, despite PPAR.alpha.'s similarity with other PPARs, exploits the

> unique structural features of PPAR.alpha.. A PPAR modulator developed using structure-assisted design would take advantage of heretofore unknown PPAR structural considerations and thus be more effective than a modulator developed using homology-based design. Potential or existent homology models cannot provide the necessary degree of specificity. A PPAR modulator designed using the structural coordinates of a crystalline form of PPAR.alpha. would also provide a starting point for the development of modulators of other PPARs. (Lambert *et al.*, paragraph 12 (emphasis added))

The above-quoted paragraph from Lambert *et al.* clearly states that once the structure of a particular protein (*e.g.*, PPAR or BACE) is solved (as it is the case in the instant application), the design of protein modulators would be greatly facilitated. As opposed to three-dimensional models generated strictly based on homology (without the benefit of the solved structure), which as quoted by the Office "cannot provide the necessary degree of specificity." Thus, the Lambert reference actually supports Applicants' position regarding the routine application of homology modeling techniques to design agents that interact with BACE, once the structure is solved.

On pages 11-13 of the Office Action, the Office alleges that, at the time of filing of the application (*i.e.*, September 22, 2000), there was "a high level of unpredictability associated with altering a polypeptide sequence with an expectation that the polypeptide will maintain the desired activity/utility" (page 12 of the Office Action). In support of this position, the Office cites to Witkoski *et al.* (*Biochemistry* 38:11643-11650) for its teaching that a single amino acid substitution can alter the specificity of beta-ketoacyl synthase.

This aspect of the rejection has been met by the claim amendments made herein. Applicants wish to clarify for the record the teaching by Witkoski *et al.* This reference provide one example of an enzyme, which is completely different from the human BACE used in the claimed methods, that shows a critical substitution for determining enzyme specificity. There are numerous other examples of proteins that are highly tolerant to amino acid replacements. For example, Bowie *et al.* (1990) *Science* 247:1306-10, page 1306, col.2, lines 12-13) cite as evidence a study carried out on the *lac* repressor. Of approximately 1500 single amino acid substitutions at 142 positions in this protein, about one-half of the substitutions were found to be "phenotypically silent:" that is, had no noticeable effect on the activity of the protein (page 1306, col. 2, lines 14-17). Presumably the other half of the substitutions exhibited effects ranging from slight to complete abolishment of repressor activity. Thus, one can expect, based on teachings by Bowie *et al.*, to find over half (and possibly well over half) of random substitutions in any given protein to result in mutated proteins with full or nearly full activity. These are far better odds than those at issue in *In re Wands,* 858 F.2d 731 (Fed. Cir. 1988), in which the Federal Circuit stated that screening many hybridomas to find the few that fell within the claims was not undue experimentation. Based on the teachings by Bowie *et al.*, one would predict that even random substitution of residues in the BACE sequence will predictably result in a majority of the mutants' having protease activity. Given the knowledge provided in the instant application about the residues interacting with the APP peptides, one of ordinary skill would have known to make changes preferentially in other regions, thereby making the predictability of success even higher than in the *lac* repressor study reported by Bowie *et al.*

In this regard, orthologs of human BACE from species as evolutionary distant as dogs, chimpanzees, rats, mice and chicken retain protease activity, while differing in amino and nucleotide sequences are shown in the Table below. This indicates that the BACE sequence can tolerate certain sequence variability, while maintaining its biological activity.

Table: Human BACE orthologs from five species

Organism	Gene	Locus	Description	Human Similarity
dog (Canis familiaris)	BACE	--	beta-site APP-cleaving enzyme	92.49*(n)* 96.95*(a)*
chimpanzee (Pan troglodytes)	BACE	--	beta-site APP-cleaving enzyme	99.72*(n)* 100*(a)*
rat (Rattus norvegicus)	BACE	--	beta-site APP cleaving enzyme	91.35*(n)* 96.41*(a)*
mouse (Mus musculus)	**BACE** BACE	9^4	beta-site APP cleaving enzyme	91.02*(n)*^{1} 96.21*(a)*^{1}
chicken (Gallus gallus)	BACE	--	beta-site APP-cleaving enzyme	82.68*(n)* 89.5*(a)*

Lastly, the Office applies the factors set forth in *In re Wands*, *supra*, to the instant specification and claims in support of the conclusion that undue experimentation would be necessary to practice the invention, as previously claimed, given that the specification discloses one crystal structure complex.

This last aspect of the rejection is also traversed. Most of these factors, *e.g.*, breadth of the claims, direction provided in the instant specification, state of the prior art have been addressed above, the substance of these arguments is reiterated here. Applicants will only address here the references, such as Branden *et al.*, cited by the Office as evidence of the high level of unpredictability in the state of the crystallography art.

As discussed above, the claims, as presently pending, specify that the three dimensional model of the BACE structure used in the claimed methods includes the relative structural coordinates of numerous (or all) residues located in the APP-binding site of human BACE, or the coordinates for its entire sequence, according to Figures 1A-1EEE, ± a root mean square deviation from the backbone atoms of BACE of no more than 1.5Å. Accordingly, *de novo* crystallization and resolution of the BACE protein is <u>not</u> required to practice the claimed invention. Newly added 42-43 include the additional step of providing a crystalline composition of BACE having the particular space group and unit cell parameters specified. In view of the claim amendments made herein, the complexities in the crystallography cited by the Office are not relevant to the present invention as claimed.

Even if the claims would require *de novo* crystallization techniques (which Applicants submit that it does not), it is submitted that such experimentation would not necessarily be undue as required to meet the lack of enablement standard set out by the CAFC in *Wands* and re-articulated in the *Falkner* decision. It is acknowledged that establishing adequate protein crystallization conditions is a tedious and time-consuming process. At least two of the references (*i.e.*, the Flower and Branden references) cited by the Office describe the availability of automated methods for speeding up "the tedious work of reproducibly setting up large numbers of crystallization experiments." *See e.g.*, Branden at page 375. Methods of producing pure and homogeneous BACE samples successful for crystallization can be readily obtained using recombinant techniques. *Id.* Even the crystallographic phase is characterized by the Flower reference as follows:

> However, even the recalcitrant discipline [crystallography phase] is yielding to the power of robotics and bioinformatics [citation omitted]. This allows many more trials to be performed and at much more accurately defined conditions than is the case for manual crystallizations. This has, in turn, led to the successful crystallization of many seemingly intractable proteins, such as several subunits from the lipocalin crustacyanin. Others have used sophisticated statistical techniques to speed the search for crystallization conditions but cutting down the number of conditions that needed to be tested. For

> example, robust multivariate statistics has been used to relate variations in experimental condition, within experimentally designed crystallization trials, to their results [citation omitted]. Although these mathematical models can not explain crystallization mechanisms, they do provide a powerful pragmatic tool allowing the setting up of crystallization trials in a more rational and more confident manner, particular when proteins are in limited supply. Flower reference at page 23.

Thus, at the Applicants' filing date, successful crystallization of "many similarly intractable proteins" had been achieved. Mathematical models for predicting crystallization trials were available to the skilled artisan. Although experimentation would have been required, it would not have been undue as specified by the enablement standard.

Applicants submit that the present application satisfies the enablement requirement as set forth in the Federal Circuit's recent *Falkner* decision, discussed above. The court held that the claims directed to poxvirus vaccine vectors, for which no example was provided, were enabled by Inglis' application by stating that:

> The Board did not err in finding Inglis' claims to be enabled as a matter of law, in light of its articulated underlying factual findings. In support of its conclusion, it noted that "there is extensive disclosure of the selection of an essential gene, its deletion or inactivation and the production of a mutated virus with said deleted or inactivated gene, albeit for herpesvirus." Moreover, because the differences between the herpesviruses and poxviruses were well known, this would have aided the person of ordinary skill in the art in her application of the lessons of the herpesvirus example in the construction of poxvirus vaccines. The Board observed that "the mere fact that the experimentation may have been difficult and time consuming does not mandate a conclusion that such experimentation would have been considered to be 'undue' in this art. Indeed, great expenditures of time and effort were ordinary in the field of vaccine preparation." (448 F. 3d at 1365; emphasis added)

In view of the foregoing arguments, the claims as presently amended, the state of the art, the guidance provided by the specification, the present specification satisfies the enablement requirement. Accordingly, reconsideration and withdrawal of the rejection of the claims, as amended herein, is respectfully requested.

Rejection of claims 12-16, 18-24, 26-27, and 33-34 under 35 U.S.C. §102(e)/103

In paragraphs 8 and 10 of the Office Action, the Office maintains the rejection of claims 12-16, 18-24, 26-27, and 33-34 under 35 U.S.C. §102(e) as being anticipated by or, in the alternative, obvious under 35 U.S.C. 103(a) over Tang et al. (U.S. Patent No. 6,545,127).

The Tang reference discloses the crystal structure of amino acids Ala14 to Thr454 of human BACE, complexed with a peptide inhibitor, OM99-2. The resulting complex forms a crystal having a space group $P2_1$, with unit cell dimensions a=53.7, b=85.9, c=109.2, with α=90.0°, β=101.4°, and γ=90.0°.

The rejection of the claims as being anticipated by the Tang reference has been met by amending independent claim 12 (and claims dependent thereon) to specify that the three dimensional structure used in the claimed methods includes the specified structural coordinates of a complex of BACE and an APP inhibitor having the sequence, SEVNStaVAEF, wherein Sta is (S)-Statine. The structure of the APP inhibitor claimed is different from the OM99-2 inhibitor disclosed by the Tang reference. In addition, the structural coordinates of the complex used in the methods as presently claimed and the complex disclosed by Tang *et al.* are different. The differences in structure between the two complexes is manifested by the different space groups and unit cell dimensions. The Tang reference discloses a crystalline complex of BACE having a space group $P2_1$, with unit cell dimensions a=53.7, b=85.9, c=109.2, with α=90.0°, β=101.4°, and γ=90.0°, whereas the instant application discloses a complex having space group I222, and unit cell parameters a=86.627, b=130.861, c=130.729, and $\alpha=\beta=\gamma=90°$. Applicants submit that the structural coordinates specified in the pending method claims are a direct function of the space group and unit cell dimensions of the BACE-APP inhibitor complex disclosed in the instant application. Therefore, claim 12 and claims dependent therefrom are not anticipated by the teachings in Tang *et al.*

Likewise, claim 20 has been amended to be directed to a method of identifying an agent using a three dimensional model having the structural coordinates of a human BACE polypeptide consisting essentially of amino acids 58-447 of SEQ ID NO:1. The BACE amino acid sequence used in the crystal complex disclosed by Tang *et al.* contains amino acids Ala14 to Thr454 of human BACE. Claim 20, as amended herein, is closed with respect to including additional residues from human BACE. It is noted for the record that the human BACE sequence recited in amended claim 20 may contain additional non-BACE elements that do not materially affect the basic and novel characteristics of the claim. Therefore, the Tang reference fails to anticipate claim 20 (and claims depending thereon) as the human BACE sequence disclosed by the crystal complexes by Tang *et al.* starts at Ala14, instead of Gly58. Similarly, as stated above, the structural coordinates of amino acids 58-447 of human BACE encompassed by the claims differ from the coordinates disclosed by Tang *et al*, as the coordinates are a direct result of the different structures of the crystal complex disclosed in the instant application compared to the crystal complex disclosed by Tang *et al.*

Moreover, newly added claims 42-43 include the additional limitation of "providing a crystalline composition of the complex, wherein the crystalline composition has space group I222, and unit cell parameters a=86.627, b=130.861, c=130.729, and α=β=γ=90°." As discussed above, the space group and unit cell dimensions recited by the new claims differ from the values of the BACE crystal complexes disclosed by Tang *et al.*

In view of the foregoing, reconsideration and withdrawal of this rejection is respectfully requested.

In paragraph 10 of the Office Action, the Office has rejected claims 12-16, 18-24, 26-27, and 33-36 under 35 U.S.C. 103(a) as being unpatentable over Tang *et al.* (US Patent 6,545,127) in view of In re Gulack 217 USPQ 401 (Fed. Cir. 1983). To support this rejection, the Office asserts that:

> [i]t would have been obvious to one of ordinary skill in the art at the time of the invention to perform rational drug design as taught by Tang et al. to identify an agent that binds to BACE and optionally including the steps of obtaining or synthesizing the compound and contacting the compound with BACE protein, wherein only non-functional descriptive material is additionally present in the claims, which do not distinguish the claimed methods from those taught by Tang et al. according to In re Gulack.

The Examiner has cited *In re Gulack* to support the proposition that the structural coordinates amount to "only non-functional descriptive material" and as such do not "distinguish the claimed methods from those taught by Tang *et al.*"

Applicants respectfully traverse the Examiner's position. The claims, as presently amended, require the use in concrete steps of specific three dimensional models having the particular structural coordinates specified that are based on the crystal structure of human BACE disclosed in the instant application, which includes, or consists essentially of, amino acids 58-447 of SEQ ID NO:1, alone or in combination with a particular BACE inhibitor having the sequence SEVNStaVAEF, wherein Sta is (S)-Statine, neither of which is taught or suggested by the Tang reference. As discussed above, the human BACE sequence used in the crystals disclosed by Tang includes amino acids Ala14 to Thr454 of human BACE, and thus includes residues 14Ala-Gly-Val-Leu-Pro-Ala-His-Gly-Thr-Gln-His-Gly-Ile-Arg-Leu-Pro-Leu-Arg-Ser-Gly-Leu-Gly-Gly-Ala-Pro-Leu-Gly-Leu-Arg-Leu-Pro-Arg-Glu-Thr-Asp-Glu-Glu-Pro-Glu-Gly-Arg-Arg-57 and 448Pro-Gln-Thr-Asp-Glu-Ser-Thr454, which are not present in the BACE three dimensional structure disclosed and claimed in the instant application. This 49-amino acid difference between the BACE sequence used in the crystals of the instant application compared to the crystals disclosed by Tang *et al.* undoubtedly leads to a new and non-obvious three dimensional structure of human BACE that would not have been expected based on the teachings by Tang *et al.*

In addition, the APP inhibitor, SEVNStaVAEF, differs from the OM99-2 disclosed by the Tang reference in several important ways: (1) it is a 9-amino acid sequence (instead of the 8-amino acid sequence of OM99-2); (2) it has an additional Ser at the N-terminus; and (3) it has the Leu-Ala isostere in OM99-2 replaced by statine-Val bond. Statine has has one extra carbon in the backbone and an extra methyl group coming off that carbon compared to the Leu residue in OM99-2. Differences in peptide length and sequence are likely to have a pronounced effect in the conformation and/or activity of the BACE-APP peptide complex. For example, as shown by the Tang reference in column 22, OM99-1 and OM99-2 have identical sequences with the exception of an additional Glutamate residue present at the N-terminus of OM99-2, which is absent in OM99-1. When comparing the inhibition of BACE using these two inhibitors OM99-1 had a Ki of $3x10^{-8}$M, whereas OM99-2 had a Ki of $9.58x10^{-9}$M (*see* Tang *et al.*, columns 28-29). Thus, slight differences in APP peptide structure can have a profound effect in the activity of BACE, and are expected to have a significant effect in the structure of a BACE-APP peptide complex. In fact, the differences in structure between the two complexes is manifested by the different space groups and unit cell dimensions of the crystal complexes disclosed by Tang compared to the complex disclosed in the present application. The Tang reference discloses a crystalline complex of BACE having a space group $P2_1$, with unit cell dimensions a=53.7, b=85.9, c=109.2, with α=90.0°, β=101.4°, and γ=90.0°, whereas the instant application discloses a complex having space group I222, and unit cell parameters a=86.627, b=130.861, c=130.729, and $\alpha=\beta=\gamma=90°$. Applicants submit that the structural coordinates specified in the pending method claims are a direct function of the space group and unit cell dimensions of the BACE-APP inhibitor complex disclosed in the instant application, thus the particular crystal structure discovered in the instant application is embedded in the claims.

The differences between the structure of BACE as encompassed by the claims (compared to the complex disclosed by Tang *et al.*) impose a change in the screening and/or design process that ultimately lead to obtaining an agent that interacts with the particular human BACE model presently claimed. Such structural information is functional, as it imparts a series of concrete steps that ultimately result in the designed or screened agent. Such new and non-obvious teachings allow for the generation of three-dimensional models for screening for agents that interact with BACE that would not have been generated based on the teachings in Tang. More particularly, the claims, as presently amended, require the use of the three-dimensional coordinates in concrete steps of (*e.g.*, performing computer fitting analysis of a candidate agent with the three dimensional model of the complex; determining the degree of association between the candidate agent and the three dimensional model of BACE; contacting the candidate agent with BACE). These steps do alter the steps performed by a computer program and/or when conducting drug screens, as explained in more detail below.

When the candidate agent is positioned in the APP-binding site of BACE, the particular structural coordinates of the BACE site recited in the claims provide a specific spatial relationship and energy surface between the binding site and the candidate agent. During the docking process, the orientation of the candidate agent is constantly adjusted in the binding site by interactive real-time energy calculations between the binding site and the candidate agent. The energy calculations provide feedback to the docking program and dictate how the computer program functions to find an energetically favorable conformation of the candidate agent. If the interaction between the binding site and the candidate agent moves uphill in energy, this feedback will dictate the computer program to resist the motion. If the interaction between the binding site and the candidate agent is favorable, the feedback with dictate the computer program to encourage the motion (*See* N. Claude-Cohen *et al.* (1990) *J. of Med. Chemistry* 33(3):883-894, submitted herewith as Exhibit G). Thus, the structural coordinates of the BACE binding site recited in the claims dictate how the computer program functions.

Furthermore, the structural coordinates of the binding site are not merely used for comparison of the structural coordinates of the candidate agent. As specified in the claims, the structural coordinates of the candidate agent are in fact changed by the structural coordinates of the BACE binding site during the docking process. As the orientation of the candidate agent is adjusted, the three-dimensional structural coordinates of the candidate agent are changed.

Therefore, the structural coordinates of the BACE binding site specified by the claimed methods impart functionality by changing the processing steps of the computer program, changing the structural coordinates of the BACE binding site and the candidate agent, which ultimately imposes a change in the screening and/or design process that leads to obtaining an agent that interacts with BACE. Such structural information is not non-functional descriptive material, as alleged by the Office, as it imparts a series of concrete steps having a functional relationship between matter and substrate. *See Gulack*, 703 F.2d at 1387.

Accordingly, reconsideration and withdrawal of the present rejection is respectfully requested.

Enclosed is a check for the Petition for Extension of Time fee. A Request for Continued Examination and appropriate fee are also being submitted herewith. Please apply any other charges or credits to deposit account 06-1050, referencing attorney docket number 16163-015001.

Respectfully submitted,

Date: October 3, 2007

Diana Collazo

Diana Collazo
Reg. No. 46,635

Fish & Richardson P.C.
225 Franklin Street
Boston, MA 02110
Telephone: (617) 542-5070
Facsimile: (617) 542-8906

21526535.doc

Exhibit A

The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor

Jie Kang, Hans-Georg Lemaire, Axel Unterbeck, J. Michael Salbaum, Colin L. Masters*, Karl-Heinz Grzeschik†, Gerd Multhaup, Konrad Beyreuther & Benno Müller-Hill

Institut für Genetik der Universität zu Köln, Weyertal 121, D-5000 Köln 41, FRG
* Neuromuscular Research Institute, Department of Pathology, University of Western Australia, Perth, 6009 Western Australia and Department of Neuropathology, Royal Perth Hospital, Perth, 6001 Western Australia
† Institut für Humangenetik, Westfälische Wilhelms Universität Münster, FRG

Alzheimer's disease[1] is characterized by a widespread functional disturbance of the human brain. Fibrillar amyloid proteins are deposited inside neurons as neurofibrillary tangles[2] and extracellularly as amyloid plaque cores[2] and in blood vessels[2]. The major protein subunit (A4) of the amyloid fibril of tangles, plaques and blood vessel deposits is an insoluble, highly aggregating small polypeptide of relative molecular mass 4,500[3-6]. The same polypeptide is also deposited in the brains of aged individuals with trisomy 21 (Down's syndrome)[3,5,6]. We have argued previously that the A4 protein is of neuronal origin and is the cleavage product of a larger precursor protein[4,6]. To identify this precursor, we have now isolated and sequenced an apparently full-length complementary DNA clone coding for the A4 polypeptide. The predicted precursor consists of 695 residues and contains features characteristic of glycosylated cell-surface receptors. This sequence, together with the localization of its gene on chromosome 21, suggests that the cerebral amyloid deposited in Alzheimer's disease and aged Down's syndrome is caused by aberrant catabolism of a cell-surface receptor.

We reasoned that the precursor of the A4 polypeptide may be produced in large amounts during the whole of life by cells of the human central nervous system. We therefore used brain tissue of a five-month-old fetus without trisomy 21 to obtain the messenger RNA needed to construct a cDNA library[7]. We used a probe designed from the N-terminal part of the A4 polypeptide to screen the library (details in Fig. 1 legend), and found several positive clones. The first one sequenced (clone 9-110) encoded the entire A4 subunit protein.

If we assume that translation starts at the first AUG, the protein would have 695 residues. It begins with a signal sequence (Figs 1 and 2) for transport through the endoplasmic reticulum membrane[8]. The signal sequence is followed by a region rich in cysteine. The last of 12 cysteines occurs at residue 187. Thus, the exact folding of these N-terminal regions may be stabilized by disulphide bridges. The next hundred residues are of peculiar composition. They include a stretch of seven uninterrupted threonine residues and are extremely rich in glutamic (28 residues) and aspartic (17 residues) acids, but contain only one lysine and two arginines. This domain would be able to bind large numbers of positive ions or polycations and may function in the electrical activity of nerve cells. The region from residue 290 to residue 597, the N-terminus of the amyloid A4 protein, contains two potential *N*-glycosylation sites at positions 467–469 and 496–498.

Like the anionic domain, the sequence of the amyloid A4 protein is also unique. We determined the complete amino-acid sequence of the A4 polypeptide isolated from both amyloid plaque cores and neurofibrillary tangles (for details see Fig. 2 legend). At most the amyloid subunit can be either 42 or 43 residues long and shows N-terminal heterogeneity as expected from previous work[3,4,6]. The 43-residue protein terminates at, and the 42-residue protein terminates just before, the first of the two threonine residues found in the potential transmembrane region (residues 625–648). The C-terminal cleavage of mature precursor during amyloidogenesis would thus have to occur within the membrane. Such a cleavage seems unlikely and the details of this process deserve further attention. The C-terminal end of the precursor is short (57 residues) compared to the cytoplasmic domain of known receptor proteins[9]. Following the putative membrane domain, there are three lysine residues (649–651) which probably interact with the phospholipid head groups in the membrane and are a characteristic feature for the junction between membrane and cytoplasmic domains of cell-surface receptors[9]. There is also a single sequence capable of asparagine-linked glycosylation (Asn-Pro-Thr). However, it is likely that this site is not glycosylated because of the presence of proline (Figs 1 and 2). The sequence and composition of the A4 precursor are not homologous to known proteins (see Fig. 1 legend) including the sequenced components of normal neurofilaments[10,11] or the PrP protein and gene isolated from the transmissible spongiform encephalopathies[12].

To determine the size of the mRNA encoding the A4 precursor we performed Northern blot hybridizations[13]. Total RNA was isolated from the same fetal cortex used for constructing the cDNA library. The RNA preparation was probed with the *Eco*RI fragment of the cloned A4 precursor cDNA (Fig. 3*a*). Two transcripts of 3.4 and 3.2 kilobases (kb) were detected. Both signals are in the size range of the cDNA insert described in Fig. 1. Similar results were obtained with mRNA from normal adult human cortex, and from cortex of sporadic Alzheimer's disease patients (data not shown).

Southern blot analyses[14] of DNA from 29 human × mouse cell hybrids[15] were used to localize human A4 precursor sequen-

Table 1 Chromosome location of human A4 precursor gene

Lane in		Human chromosomes retained																								Hybridization
																										(8.8-kb band)
Fig. 3*b*	Hybrid clone	1	2	3	4	5	6	7	8	9	10	11	12	13.	14	15	16	17	18	19	20	21	22	X	Y	in Fig. 3*b*
1	Human genome	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	+
2	RAG-SU 3-1-2-1			O		X	O	O														X		X		+
3	RAG RU 4-13	O		X		O	X	O			O				O			X		O		X	X	O	X	+
4	RAG 194-7	X		O	X	X			X		X		O							X		X		O		+
5	A9-SU 1-2		X	O			X							X	X				X			X				+
6	MS2 B82-1a-141		X	X	X			X	X			X		X		X	X	X	X	X		X	X	X	X	+
7	RAG ANLY 1	X		X	X	X	X						X		O				X			X				+
8	RAG 194-5-5			O	X	X	X	X			X		X	X	X		X				X	X	X	O		+
9	RAG GO-4			X	X			X	X					X	X	X			X				X	O		−
10	RAG PI 7-2	X		X		X	X	X	X	X		O	X		X	O			X				X	O		−
11	RAG 610-4-5-1	O		X	X	X	O						X	O	X	X			O		X		X	X		−
12	RAG mouse																									−

X, whole chromosome; O, chromosome fragment.

```
                                  AGTTTCCTCGGCAGCGGTAGGCGAGA  -121

GCACGCGGAGGAGCGTGCGCGGGGCCCCGGGAGACGGCGGCGGTGGCGGCGCGGGCAGAG   -61

CAAGGACGCGGCGGATCCCACTCGCACAGCAGCGCACTCGGTGCCCCGCGCAGGGTCGCG    -1

ATGCTGCCCGGTTTGGCACTGCTCCTGCTGGCCGCCTGGACGGCTCGGGCGCTGGAGGTA    60
M  L  P  G  L  A  L  L  L  L  A  A  W  T  A  R  A  L  E  V
1                           10                            20

CCCACTGATGGTAATGCTGGCCTGCTGGCTGAACCCCAGATTGCCATGTTCTGTGGCAGA   120
P  T  D  G  N  A  G  L  L  A  E  P  Q  I  A  M  F  C  G  R
                           30                            40

CTGAACATGCACATGAATGTCCAGAATGGGAAGTGGGATTCAGATCCATCAGGGACCAAA   180
L  N  M  H  M  N  V  Q  N  G  K  W  D  S  D  P  S  G  T  K
                           50                            60

ACCTGCATTGATACCAAGGAAGGCATCCTGCAGTATTGCCAAGAAGTCTACCCTGAACTG   240
T  C  I  D  T  K  E  G  I  L  Q  Y  C  Q  E  V  Y  P  E  L
                           70                            80

CAGATCACCAATGTGGTAGAAGCCAACCAACCAGTGACCATCCAGAACTGGTGCAAGCGG   300
Q  I  T  N  V  V  E  A  N  Q  P  V  T  I  Q  N  W  C  K  R
                           90                            100

GGCCGCAAGCAGTGCAAGACCCATCCCCACTTTGTGATTCCCTACCGCTGCTTAGTTGGT   360
G  R  K  Q  C  K  T  H  P  H  F  V  I  P  Y  R  C  L  V  G
                           110                           120

GAGTTTGTAAGTGATGCCCTTCTCGTTCCTGACAAGTGCAAATTCTTACACCAGGAGAGG   420
E  F  V  S  D  A  L  L  V  P  D  K  C  K  F  L  H  Q  E  R
                           130                           140

ATGGATGTTTGCGAAACTCATCTTCACTGGCACACCGTCGCCAAAGAGACATGCAGTGAG   480
M  D  V  C  E  T  H  L  H  W  H  T  V  A  K  E  T  C  S  E
                           150                           160

AAGAGTACCAACTTGCATGACTACGGCATGTTGCTGCCCTGCGGAATTGACAAGTTCCGA   540
K  S  T  N  L  H  D  Y  G  M  L  L  P  C  G  I  D  K  F  R
                           170                           180

GGGGTAGAGTTTGTGTGTTGCCCACTGGCTGAAGAAAGTGACAATGTGGATTCTGCTGAT   600
G  V  E  F  V  C  C  P  L  A  E  E  S  D  N  V  D  S  A  D
                           190                           200

GCGGAGGAGGATGACTCGGATGTCTGGTGGGGCGGAGCAGACACAGACTATGCAGATGGG   660
A  E  E  D  D  S  D  V  W  W  G  G  A  D  T  D  Y  A  D  G
                           210                           220

AGTGAAGACAAAGTAGTAGAAGTAGCAGAGGAGGAAGAAGTGGCTGAGGTGGAAGAAGAA   720
S  E  D  K  V  V  E  V  A  E  E  E  E  V  A  E  V  E  E  E
                           230                           240

GAAGCCGATGATGACGAGGACGATGAGGATGGTGATGAGGTAGAGGAAGAGGCTGAGGAA   780
E  A  D  D  D  E  D  D  E  D  G  D  E  V  E  E  E  A  E  E
                           250                           260

CCCTACGAAGAAGCCACAGAGAGAACCACCAGCATTGCCACCACCACCACCACCACCACA   840
P  Y  E  E  A  T  E  R  T  T  S  I  A  T  T  T  T  T  T  T
                           270                           280

GAGTCTGTGGAAGAGGTGGTTCGAGTTCCTACAACAGCAGCCAGTACCCCTGATGCCGTT   900
E  S  V  E  E  V  V  R  V  P  T  T  A  A  S  T  P  D  A  V
                           290                           300

GACAAGTATCTCGAGACACCTGGGGATGAGAATGAACATGCCCATTTCCAGAAAGCCAAA   960
D  K  Y  L  E  T  P  G  D  E  N  E  H  A  H  F  Q  K  A  K
                           310                           320

GAGAGGCTTGAGGCCAAGCACCGAGAGAGAATGTCCCAGGTCATGAGAGAATGGGAAGAG  1020
E  R  L  E  A  K  H  R  E  R  M  S  Q  V  M  R  E  W  E  E
                           330                           340

GCAGAACGTCAAGCAAAGAACTTGCCTAAAGCTGATAAGAAGGCAGTTATCCAGCATTTC  1080
A  E  R  Q  A  K  N  L  P  K  A  D  K  K  A  V  I  Q  H  F
                           350                           360

CAGGAGAAAGTGGAATCTTTGGAACAGGAAGCAGCCAACGAGAGACAGCAGCTGGTGGAG  1140
Q  E  K  V  E  S  L  E  Q  E  A  A  N  E  R  Q  Q  L  V  E
                           370                           380

ACACACATGGCCAGAGTGGAAGCCATGCTCAATGACCGCCGCCGCCTGGCCCTGGAGAAC  1200
T  H  M  A  R  V  E  A  M  L  N  D  R  R  R  L  A  L  E  N
                           390                           400

TACATCACCGCTCTGCAGGCTGTTCCTCCTCGGCCTCGTCACGTGTTCAATATGCTAAAG  1260
Y  I  T  A  L  Q  A  V  P  P  R  P  R  H  V  F  N  M  L  K
                           410                           420

AAGTATGTCCGCGCAGAACAGAAGGACAGACAGCACACCCTAAAGCATTTCGAGCATGTG  1320
K  Y  V  R  A  E  Q  K  D  R  Q  H  T  L  K  H  F  E  H  V
                           430                           440

CGCATGGTGGATCCCAAGAAAGCCGCTCAGATCCGGTCCCAGGTTATGACACACCTCCGT  1380
R  M  V  D  P  K  K  A  A  Q  I  R  S  Q  V  M  T  H  L  R
                           450                           460

GTGATTTATGAGCGCATGAATCAGTCTCTCTCCCTGCTCTACAACGTGCCTGCAGTGGCC  1440
V  I  Y  E  R  M  N  Q  S  L  S  L  L  Y  N  V  P  A  V  A
                           470                           480

GAGGAGATTCAGGATGAAGTTGATGAGCTGCTTCAGAAAGAGCAAAACTATTCAGATGAC  1500
E  E  I  Q  D  E  V  D  E  L  L  Q  K  E  Q  N  Y  S  D  D
                           490                           500

GTCTTGGCCAACATGATTAGTGAACCAAGGATCAGTTACGGAAACGATGCTCTCATGCCA  1560
V  L  A  N  M  I  S  E  P  R  I  S  Y  G  N  D  A  L  M  P
                           510                           520

TCTTTGACCGAAACGAAAACCACCGTGGAGCTCCTTCCCGTGAATGGAGAGTTCAGCCTG  1620
S  L  T  E  T  K  T  T  V  E  L  L  P  V  N  G  E  F  S  L
                           530                           540

GACGATCTCCAGCCGTGGCATTCTTTTGGGGCTGACTCTGTGCCAGCCAACACAGAAAAC  1680
D  D  L  Q  P  W  H  S  F  G  A  D  S  V  P  A  N  T  E  N
                           550                           560

GAAGTTGAGCCTGTTGATGCCCGCCCTGCTGCCGACCGAGGACTGACCACTCGACCAGGT  1740
E  V  E  P  V  D  A  R  P  A  A  D  R  G  L  T  T  R  P  G
                           570                           580

TCTGGGTTGACAAATATCAAGACGGAGGAGATCTCTGAAGTGAAGATGGATGCAGAATTC  1800
S  G  L  T  N  I  K  T  E  E  I  S  E  V  K  M  D  A  E  F
                           590                           600

CGACATGACTCAGGATATGAAGTTCATCATCAAAAATTGGTGTTCTTTGCAGAAGATGTG  1860
R  H  D  S  G  Y  E  V  H  H  Q  K  L  V  F  F  A  E  D  V
                           610                           620

GGTTCAAACAAAGGTGCAATCATTGGACTCATGGTGGGCGGTGTTGTCATAGCGACAGTG  1920
G  S  N  K  G  A  I  I  G  L  M  V  G  G  V  V  I  A  T  V
                           630                           640

ATCGTCATCACCTTGGTGATGCTGAAGAAGAAACAGTACACATCCATTCATCATGGTGTG  1980
I  V  I  T  L  V  M  L  K  K  K  Q  Y  T  S  I  H  H  G  V
                           650                           660

GTGGAGGTTGACGCCGCTGTCACCCCAGAGGAGCGCCACCTGTCCAAGATGCAGCAGAAC  2040
V  E  V  D  A  A  V  T  P  E  E  R  H  L  S  K  M  Q  Q  N
                           670                           680

GGCTACGAAAATCCAACCTACAAGTTCTTTGAGCAGATGCAGAACTAGACCCCCGCCACA  2100
G  Y  E  N  P  T  Y  K  F  F  E  Q  M  Q  N  *
                           690

GCAGCCTCTGAAGTTGGACAGCAAAACCATTGCTTCACTACCCATCGGTGTCCATTTATA  2160

GAATAATGTGGGAAGAAACAAACCCGTTTTATGATTTACTCATTATCGCCTTTTGACAGC  2220

TGTGCTGTAACACAAGTAGATGCCTGAACTTGAATTAATCCACACATCAGTAATGTATTC  2280

TATCTCTCTTTACATTTTGGTCTCTATACTACATTATTAATGGGTTTTGTGTACTGTAAA  2340

GAATTTAGCTGTATCAAACTAGTGCATGAATAGATTCTCTCCTGATTATTTATCACATAG  2400

CCCCTTAGCCAGTTGTATATTATTCTTGTGGTTTGTGACCCAATTAAGTCCTACTTTACA  2460

TATGCTTTAAGAATCGATGGGGGATGCTTCATGTGAACGTGGGAGTTCAGCTGCTTCTCT  2520

TGCCTAAGTATTCCTTTCCTGATCACTATGCATTTTAAAGTTAAACATTTTTAAGTATTT  2580

CAGATGCTTTAGAGAGATTTTTTTTCCATGACTGCATTTTACTGTACAGATTGCTGCTTC  2640

TGCTATATTTGTGATATAGGAATTAAGAGGATACACACGTTTGTTTCTTCGTGCCTGTTT  2709

TATGTGCACACATTAGGCATTGAGACTTCAAGCTTTTCTTTTTTTGTCCACGTATCTTTG  2760

GGTCTTTGATAAAGAAAAGAATCCCTGTTCATTGTAAGCACTTTTACGGGGCGGGTGGGG  2820

AGGGGTGCTCTGCTGGTCTTCAATTACCAAGAATTCTCCAAAACAATTTTCTGCAGGATG  2880

ATTGTACAGAATCATTGCTTATGACATGATCGCTTTCTACACTGTATTACATAAATAAAT  2940

TAAATAAAATAACCCCGGGCAAGACTTTTCTTTGAAGGATGACTACAGACATTAAATAAT  3000

CGAAGTAATTTTGGGTGGGGAGAAGAGGCAGATTCAATTTTCTTTAACCAGTCTGAAGTT  3060

TCATTTATGATACAAAAGAAGATGAAAATGGAAGTGGCAATATAAGGGGATGAGGAAGGC  3120

ATGCCTGGACAAACCCTTCTTTTAAGATGTGTCTTCAATTTGTATAAAATGGTGTTTTCA  3180

TGTAAATAAATACATTCTTGGAGGAGC-poly(A) tail
```

Fig. 1 Nucleotide sequence and predicted amino-acid sequence of a cDNA clone encoding the precursor of the amyloid A4 protein of Alzheimer's disease. Nucleotide residues are numbered in the 5′ to 3′ direction, beginning with the first base of the initiation codon AUG. The untranslated sequence directly following the poly(G) tail is indicated by negative numbers. The sequence shows a 695-residue open reading frame. The deduced amino-acid sequence is numbered, the amino-terminal methionine included. The amino-acid sequence of the A4 polypeptide (includes data from refs 3 and 4) is boxed. Zigzag underline, a probable membrane-spanning sequence of the A4 polypeptide. The synthetic oligonucleotide mixture used as probe is indicated as a line above the corresponding cDNA. Polyadenylation signals are underlined. Nucleotide 3207 is followed by a poly(dA) tail linked to the vector DNA. A computer search for homology was unsuccessful using computer programs of the University of Wisconsin Genetics Computer Group (UWGCG). The entire cDNA insert includes 3,353 bp and *Eco*RI sites at bp positions 1,795 and 2,851. The amino-acid composition of the 695-residue A4 precursor is A57, C12, D47, E85, F17, G31, H25, I23, K38, L52, M21, N28, P31, Q33, R33, S30, T45, V62, W8, Y17 resulting a calculated M_r of 78,644.45.

Methods. Total RNA was extracted[24] from cortex of the brain of a 5-month-old aborted human fetus. Poly(A)$^+$ RNA was isolated by oligo(dT)-cellulose chromatography[25]. A cDNA library was constructed by the method of Okayama and Berg[7] with 3 µg poly(A)$^+$ RNA and 1.5 µg vector-primer DNA. The cDNA was transformed into *Escherichia coli* HB101. Transformants (10^5) were screened[26] with a probe consisting of a mixture of 64 20-mers, $5'\mathrm{TT}^{\mathrm{T}}_{\mathrm{C}}\mathrm{TG}^{\mathrm{A}}_{\mathrm{G}}\mathrm{TG}^{\mathrm{A}}_{\mathrm{G}}\mathrm{TGCAC}^{\mathrm{T}}_{\mathrm{C}}\mathrm{T}^{\mathrm{CA}}_{\mathrm{GG}}\mathrm{TA}3'$, deduced from residues 10–16 of the A4 polypeptide. Of the nine positive clones obtained, one (clone 9–110) was sequenced by the dideoxy chain termination method[27] on both strands with the help of synthetic primers. The primers were synthesized on a DNA-synthesizer (Applied Biosytems)[28]. For sequence comparison, UWGCG programs Wordsearch and Gap were applied to the data libraries of EMBL and NBR7.

a

A Asp-Ala-Glu-Phe-Arg-His-Asp-Ser-Gly-Tyr-
1 5 10
Glu-Val-His-His-Gln-Lys-Leu-Val-Phe-Phe
11 15 20
Ala-Glu-Asp-Val-Gly-Ser-Asn-Lys-Gly-Ala
21 25 30
Ile-Xxx-Xxx-Leu-Met-Val
31 35 36

B Gly-Ala-Ile-Xxx-Xxx-Leu-Met-Val
29 36

C Val-Gly-Gly-Val-Val-Ile-Ala
36 42

D Val-Gly-Gly-Val-Val-Ile-Ala-Thr
36 43

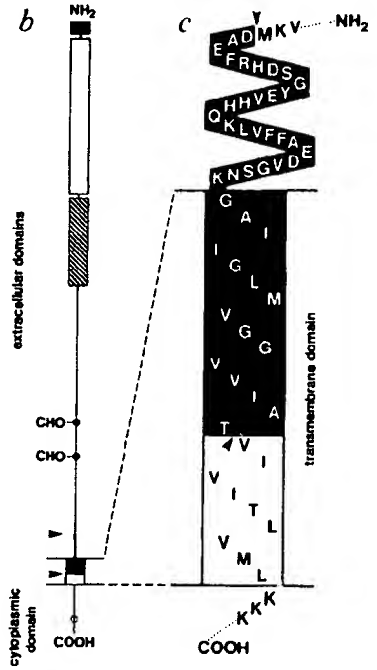

Fig. 2 Model of A4 precursor as a cell-surface glycoprotein. *a*, Amino-acid sequence of amyloid plaque core-A4 polypeptide. Amyloid from plaque cores was purified from human post-mortem brain homogenates[3,6]. The amino-terminal sequence shown (A) was obtained for amyloid plaque core-A4 polypeptides. To determine the internal and the C-terminal portion of the A4 polypeptide, the corresponding trypsin digestion (B) and cyanogen bromide cleavage products (C, D) were sequenced. However, we were unable to assign residues 32 and 33 unambiguously. Amyloid plaque core preparations from different Alzheimer brains suggest that the amyloid protein may consist variably of at most 42 or 43 residues (C, D). *b*, Proposed domain structure: black box, signal sequence[8], open box, cysteine-rich region; hatched box, highly negatively charged domain (45% Asp and Glu residues); filled circles, *N*-glycosylation sites. The A4 amyloid is shown in black between the two arrowheads and reaches into the (boxed) hydrophobic transmembrane segment. The putative cytoplasmic domain includes the sequence Asn-Pro-Thr (open circle). *c*, The full-length A4 amyloid protein (42–43 residues) as deposited in the brain of patients with Alzheimer's disease (shaded region), includes 28 residues of the extracellular domain and 14–15 residues of the transmembrane domain of the A4 precursor. The group of three Lys residues following the 24 hydrophobic amino acids of the putative transmembrane domain are typical for the junction between membrane and cytoplasmic domains of cell surface receptors[9].

Methods. The amyloid plaque cores were extracted with formic acid as described[3,4] and the formic acid-soluble protein was applied to the gas-phase protein sequencer (Model 470 A, Applied Biosystems) described by Hewick *et al.*[29]. The amino-acid derivatives released were determined by reverse-phase HPLC in on-line configuration using 40% of the sample (Model 120A, Applied Biosystems). Four different amyloid plaque core preparations were sequenced; typically, 400 pmol of total N-terminal derivatives was obtained from 900 pmol of protein (estimated by quantitative amino-acid analysis). For the tryptic digest in *b*, 10 pmol of sequence was obtained from the estimated 5 nmol of acetylated protein; 20 pmol (*c*) and 50 pmol (*d*) of sequence was obtained for the cyanogen bromide cleavages from an estimated 2 nmol of acetylated protein (details to be published elsewhere).

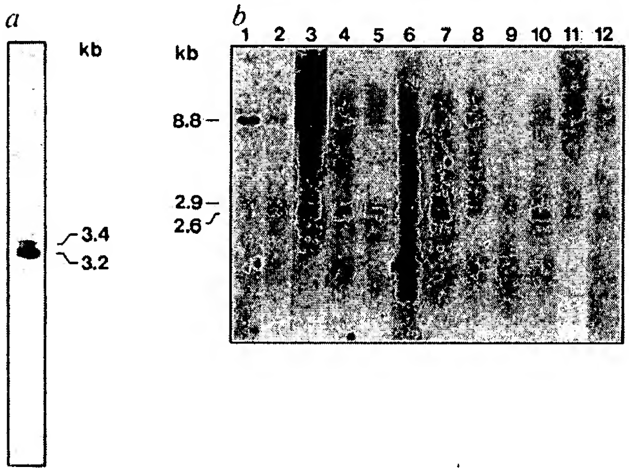

Fig. 3 Hybridization analyses. *a*, Identification of A4 precursor mRNA by Northern blot analysis[13]. RNA was isolated from human embryonic cortex as in Fig. 1. RNA (25 μg) was heated for 15 min at 65 °C in the presence of 50% formamide, subjected to electrophoresis on a 1.6% agarose gel containing 6 M urea/25 mM sodium citrate, *p*H 3.5[30] and transferred to Zetabind membrane (AMF Cuno). Blots were hybridized at 65 °C with nick-translated[31] 1,056-bp *Eco*RI fragment of clone 9–110 (bp positions 1,795–2,851, Fig. 1) in 6 × SSC, 50 mM sodium phosphate *p*H 6.5, 5 × Denhardt's, 0.5% SDS, 1 mM EDTA, 100 μg ml^{-1} denatured calf thymus DNA. The filters were washed to a final stringency of 0.2 × SSC, 0.1% SDS at 65 °C. Autoradiography was for 24 h at −70 °C using intensifying screens. An RNA ladder (BRL) was used as size marker. *b*, Chromosome mapping of A4 precursor locus. The 1,056-bp cDNA *Eco*RI fragment of clone 9–110 (Fig. 1) was used for Southern blot analysis[14] of *Eco*RI-digested human × mouse RAG or mouse A9 hybrid cell line DNAs. DNA from the somatic cell hybrids was purified from isolated nuclei as described in detail elsewhere[15]. The *Eco*RI-digested DNA (~10 μg) from each of 29 hybrid cell samples was electrophoresed through a 0.8% agarose gel and blotted onto nylon (Pall). Results from 10 of the 29 hybrids analysed are presented here and in Table 1. Hybridization and washing were as in *a*. Lane 1, human genomic DNA; lanes 2–8, hybrid cell lines containing human chromosome 21; lanes 9–11, hybrid cell lines missing human chromosome 21; lane 12, mouse RAG genomic DNA. The chromosome distribution is given in Table 1.

ces to chromosome 21. The ^{32}P-labelled *Eco*RI fragment of the human A4 precursor cDNA was hybridized to Southern filters of *Eco*RI-digested DNA from control tissue and hybrid cell lines. Two human-specific fragments of 8.8 and 2.9 kb and a mouse-specific fragment of 2.6 kb were observed (Fig. 3*b*). The 8.8 kb fragment was used to score for the presence of the human A4 precursor gene in the DNA from the hybrid cell lines. Chromosome 21 was the only human chromosome that showed perfect concordance with human A4 precursor sequences (Fig. 3*b*, lanes 2–8, and Table 1). Every other human chromosome could be ruled out by at least two discordant hybrids.

It may be significant that chromosome 21 is the chromosome that is duplicated in Down's syndrome. We have evidence that the A4 gene is close to the part of the chromosome that is duplicated (unpublished results). There could be a more complex interaction between other chromosome 21 gene products and the biosynthesis and processing of the A4 precursor in Down's syndrome, rather than a simple gene-dosage effect.

The transmissible spongiform encephalopathies, scrapie, CJD and kuru[16] are progressive degenerative disorders of the nervous sytem sharing many clinical and pathological features with Alzheimer's disease. In these unusual infectious diseases, an abnormally post-translationally modified protein (PrP)[12] accumulates in the form of amyloid structures in the brain[17-20]. However, the corresponding gene for the PrP protein precursor has been localized to chromosome 20[21]. Nevertheless, we and Oesch *et al.*[12] show that the two proteins share many similarities, including features characteristic of glycosylated membrane receptors. It is possible that the putative membrane-spanning domains of both proteins share an amyloid-forming or amyloid-inducing potential.

The availability of a complete A4 precursor cDNA now permits a systematic study of the normal and abnormal biology of this amyloidogenic protein in Down's syndrome, familial and sporadic Alzheimer's disease and related disorders. During the preparation of this manuscript we have learnt that two other groups have isolated cDNA clones coding for A4[22,23].

We thank Robert G. Rohwer and Caroline Hilbich for comments, Klaus Olek for a gift of Okayama-Berg primer, Thomas Herget for RNA, and Brigitte Kisters and Pauline Klein for help with the computers. This work was supported by grants from the Deutsche Forschungsgemeinschaft through SFB 74 A1, A2 and GR 373, and the National Health and Medical Research Council of Australia. The sequence data in this publication have been submitted to the EMBL/GenBank Libraries under the accession number Y00264.

Received 22 December 1986; accepted 23 January 1987.

1. Alzheimer, A. *Allg. Z. Psychiat.* **64,** 146–148 (1907).
2. Katzman, R. (ed.) *Biological Aspects of Alzheimer's Disease* (*Banbury Report 15*) (Cold Spring Harbor Laboratory, New York, 1983).
3. Masters, C. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **82,** 4245–4249 (1985).
4. Masters, C. L. *et al. EMBO J.* **4,** 2757–2763 (1985).
5. Glenner, G. G. & Wong, C. W. *Biochem. biophys. Res. Commun.* **120,** 1131–1135 (1984).
6. Beyreuther, K. *et al.* in *Discussions in Neuroscience* Vol. 3 (eds Bignami, A., Bolis, L. & Gajdusek, D. C. 68–79 Fondation pour l'Étude du Système Nerveux, Geneva (1986).
7. Okayama, H. & Berg, P. *Molec. cell. Biol.* **3,** 280–289 (1983).
8. Blobel, G. & Dobberstein, B. *J. Cell Biol.* **67,** 852–862 (1975).
9. Yarden, Y. *et al. Nature* **323,** 226–232 (1986).
10. Geisler, N., Plessmann, U. & Weber, K. *FEBS Lett.* **182,** 475–478 (1985).
11. Lewis, S. A. & Cowman, N. J. *Molec. cell. Biol.* **6,** 1529–1534 (1986).
12. Oesch, B. *et al. Cell* **40,** 735–746 (1985).
13. Thomas, P. S. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5201–5205 (1980).
14. Southern, E. M. *J. molec. Biol.* **98,** 503–517 (1975).
15. Lötscher, E. *et al. Nature* **320,** 456–458 (1986).
16. Gajdusek, D. C. *Science* **197,** 943–960 (1977).
17. Multhaup, G. *et al. EMBO J.* **4,** 1495–1501 (1985).
18. Merz, P. A., Sommerville, R. A., Wisniewski, H. M. & Iqbal, K. *Acta neuropath.* **54,** 63–74 (1981).
19. Diringer, H. *et al. Nature* **306,** 476–478 (1983).
20. Prusiner, S. B. *et al. Cell* **35,** 349–358 (1983).
21. Liao, Y-C. J., Lebo, R. V., Clawson, G. A. & Smuckler, E. A. *Science* **233,** 364–367 (1986).
22. Goldgaber, D., Lerman, M. I., McBride, W. O., Sallow, V. & Gajdusek, D. C. *Science* (in the press).
23. Robakis, N. K. *et al. Lancet* (manuscript in preparation).
24. Bernstein, S. L., Gioco, A. E. & Kaplan, B. B. *J. Neurogenet.* **1,** 71–86 (1983).
25. Aviv, H. & Leder, P. *Proc. natn. Acad. Sci. U.S.A.* **69,** 1408–1412 (1972).
26. Woods, D. E., Markham, A. F., Ricker, A. T., Goldberger, G. & Colten, H. R. *Proc. natn. Acad. Sci. U.S.A.* **79,** 5661–5665 (1982).
27. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
28. Matteucci, M. D. & Caruthers, M. H. *J. Am. Chem. Soc.* **103,** 3185–3190 (1981).
29. Hewick, R. M., Hunkapiller, M. W., Hood, L. E. & Dreyer, W. J. *J. biol. Chem.* **256,** 7990–7997 (1981).
30. Lehrach, H., Diamond, D., Wozney, J. M. & Boedtker, H. *Biochemistry* **16,** 4743–4751 (1977).
31. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. *J. molec. Biol.* **113,** 237–251 (1977).

GUIDE TO AUTHORS

Please follow these guidelines so that your manuscript may be handled expeditiously.
Nature is an international journal covering all the sciences. Contributors should therefore bear in mind those readers for whom English is a second language and those who work in other fields. Please write clearly and simply, avoiding unnecessary technical terminology.
Nature's staff will edit manuscripts to those ends if necessary. Contributors should check their galley proofs carefully.
Because of the competition for space, many of the papers submitted for publication cannot be accepted. For this reason, and because brevity is a great assistance to readers, papers should be as brief as is consistent with intelligibility. Please note that one printed page of *Nature*, without diagrams or other interruptions of the text, has fewer than 1,300 words.

CATEGORIES OF PAPER

Review Articles survey recent developments in a field. Most are commissioned, but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews coordinator. Length is negotiable in advance but must not exceed six pages of *Nature*.
Articles are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text or more than six display items (figures and tables) and should not occupy more than five pages of *Nature*.
Articles should be accompanied by a heading of 50–80 words written to advertise their contents in general terms, to which editors will pay particular attention. A heading (printed in italic type) is not an abstract and should not usually contain numbers or measurements. The study should be more carefully introduced in the first two or three paragraphs, which should also briefly summarize its results and implications.
Articles may contain a few subheadings of two or three words. The meaning of the text should not depend on the subheadings, whose function is to break up the text and to point to what follows. There should be fewer than 50 references.
Letters to Nature are short reports of an outstanding novel finding whose implications are general and important enough to be of interest to those outside the field. Letters should not have more than 1,000 words of text or more than four display items and should not occupy more than two pages of *Nature*. The first paragraph should describe, in not more than 150 words, the origins and chief conclusions of the study. Letters do not have subheadings or more than 30 references.
Commentary articles deal with issues in, or arising from, research that are also of interest to readers outside research. Some are commissioned, most are unsolicited. They are normally between one and four pages of *Nature* in length.
News and Views articles are intended to inform non-specialist readers about a recently published advance. Suggestions should be made to the News and Views coordinator. Illustrations are welcome. Proposals for meeting reports should be agreed in advance.
Scientific Correspondence is for the discussion of scientific matters, including contributions published in *Nature*. Priority is given to contributions of less than 500 words and five references. Figures are welcome.

GENERAL

All contributions submitted for publication in *Nature* should conform with these rules.
Manuscripts should be typed, double-spaced, with a good-quality printer, on one side of the paper only. Four copies are required, each accompanied by lettered artwork. Four copies of half-tones should be included. Reference lists, figure legends and so on should be on separate sheets, all of which should be double-spaced and numbered. Copies of relevant manuscripts in press or submitted for publication elsewhere should be included, clearly marked as such. Revised and resubmitted manuscripts should also be clearly marked as such and labelled with their reference numbers.
Titles should say what the paper is about with the minimum of technical terminology and in fewer than 80 characters. Authors should avoid active verbs, numerical values, abbreviations and punctuation and should include one or two key words for indexing purposes.
Artwork should be marked individually and clearly with the author's name, figure number and, when known, manuscript reference number. Original artwork should be unlettered. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are permitted only if the parts are closely related, either experimentally or logically. Suggestions for cover illustrations, with captions, are welcome. Original artwork (and one copy of the manuscript) will be returned when manuscripts cannot be published.
Figure legends must not exceed 300 words and ideally should be much shorter. Use telegraphic form wherever possible. First describe the figure then, briefly, the method. Reference to a method described elsewhere is preferable to a full description. Methods should not be described in detail in the text.
References should be numbered sequentially as they appear in the text, followed by those in tables and finally those in the figure legends. Reference numbers apply only to papers published or in the press, and a different number should be given to each paper cited. All other forms of reference (including unrefereed abstracts) should be included in the text as personal communication, manuscript in preparation or preprint (with number and institution where appropriate). Text should not be included in the reference list. References should be abbreviated according to the *World List of Scientific Periodicals* (fourth edition, Butterworth, London, 1963-1965). The first and last page numbers should be cited. Reference to books should clearly indicate the publisher and the date and place of publication.
Abbreviations, symbols, units and Greek letters should be identified the first time they are used. Acronyms should be avoided as much as possible and, when used, defined. Editors will shorten words if necessary. In case of doubt, SI units should be used.
Footnotes should not be used except for changed addresses or to identify the corresponding author if different from the first-named.
Acknowledgements should be brief. Grant numbers and contribution numbers are not allowed.
Submission. *All manuscripts may be submitted either to London or Washington. They should not be addressed to editors by name. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value $5' to prevent the imposition of import duty and value-added tax (VAT).*

Exhibit B

Volume 17 Number 2 1989 Nucleic Acids Research

The PreA4$_{695}$ precursor protein of Alzheimer's disease A4 amyloid is encoded by 16 exons

H.G.Lemaire, J.M.Salbaum[1], G.Multhaup[1], J.Kang, R.M.Bayney[2], A.Unterbeck[2], K.Beyreuther[1] and B.Müller-Hill

Institute für Genetik der Universität zu Köln, Weyertal 121, D-5000 Köln 41, [1]Zentrum für Molekulare Biologie, Universität Heidelberg, Im Neuenheimer Feld 282, D-6900 Heidelberg, FRG and [2]Molecular Therapeutics Inc., 400 Morgan Lane, West Haven, CT 06516, USA

Received November 17, 1988; Revised and Accepted December 22, 1988 Accession no. X13428

ABSTRACT

Alzheimer's disease (AD) is characterized by the cerebral deposition of fibrillar aggregates of the amyloid A4 protein. Complementary DNA's coding for the precursor of the amyloid A4 protein have been described. In order to identify the structure of the precursor gene relevant clones from several human genomic libraries were isolated. Sequence analysis of the various clones revealed 16 exons to encode the 695 residue precursor protein (PreA4$_{695}$) of Alzheimer's disease amyloid A4 protein. The DNA sequence coding for the amyloid A4 protein is interrupted by an intron. This finding supports the idea that amyloid A4 protein arises by incomplete proteolysis of a larger precursor, and not by aberrant splicing.

INTRODUCTION

Alzheimer's disease (1) is the most common cause of dementia, afflicting about two million people in the USA (2). It is characterized by the formation of intraneuronal neurofibrillary tangles (3,4,5), extracellular amyloid plaques (3,4,5) and cerebrovascular amyloid deposits (5,6) in the brain. The major constituent of these depositions is the amyloid A4 protein or β-protein (3,4,6).

Recently, we isolated and sequenced a full-length cDNA clone encoding the fetal brain precursor of the amyloid A4 protein and localized the gene (PAD gene (7)) on chromosome 21 (8). The structure of the deduced amino acid sequence suggests that the fetal brain PreA4$_{695}$ protein is a glycosylated cell-surface receptor consisting of an N-terminal signal sequence, three extracellular domains, a transmembrane region and a cytoplasmic domain. The membrane spanning domain corresponds to residues 625–648 of the PreA4$_{695}$ protein and overlaps with the amyloid A4 peptide sequence (597–639). Three other groups reported the finding of longer transcripts of the PAD gene which all contain an extra exon encoding a peptide that is very similar to the Kunitz family of protease inhibitors (9,10,11). Schubert et al. (12) showed residues 18–44 of the amyloid A4 precursor protein to be very similar to a heparan sulfate proteoglycan core protein found in the nerve cell line PC12. Here we report exon–intron boundaries of the PAD gene. Our work excludes the possibility that the amyloid A4 peptide could be the product of alternative splicing.

MATERIALS AND METHODS

Genomic libraries and screening conditions

Four different libraries were used: a) a chromosome 21 library (*Hind*III fragments in *lambda* charon 21A, courtesy of Dr M. Van Dilla) which was constructed at the Lawrence Livermore National Laboratory, Livermore, CA, under the auspices of the National Laboratory Gene Library Project, sponsored by the US Department of Energy, b) a

```
1                                                       57
|                                                       |
ATGCTGCCCGGTTTGGCACTGCTCCTGCTGGCCGCCTGGACGGCTCGGGCGCTGGAGgtg
MetLeuProGlyLeuAlaLeuLeuLeuLeuAlaAlaTrpThrAlaArgAlaLeuGlu
1                                                     19

ggtgccgcgcctcggaaggggggggaggctgcacggtggggacgcgatacccccccaatac

cttaacccaagtctttaatgcagagaagccgggggtccgtcaatgggacccctctcct..

........tggcacaaatattattttcagtcttggtcatcaggcagggcctggcaggcat

catgctctgtgtgagtcagtggagtttattataagaatgcggtagcctccacagtgatta

tatttattataggaaaacgtgtaattttgttgtcagaacctttcagcaggtaccacaaaa

tctcatatgggaaaacgtgtaaatttttcatgaataaattctttcggtattggtaatttc

cttctctgttttagtataaaagaaagcattgccttataggtggctaagaacatctgtagc

atacataacacttaatctaaaggagtgttgaagaccgggctgattcctaattgagaatga

gcaagaatagaactctttttgatagttatccctgttcttcctccaagcctctgccttgga

                                       58
                                       |
gctatggatactataactaactgaagcttcttctttcagGTACCCACTGATGGTAATGCT
                                       ValProThrAspGlyAsnAla
                                       20

GGCCTGCTGGCTGAACCCCAGATTGCCATGTTCTGTGGCAGACTGAACATGCACATGAAT
GlyLeuLeuAlaGluProGlnIleAlaMetPheCysGlyArgLeuAsnMetHisMetAsn

GTCCAGAATGGGAAGTGGGATTCAGATCCATCAGGGACCAAAACCTGCATTGATACCAAG
ValGlnAsnGlyLysTrpAspSerAspProSerGlyThrLysThrCysIleAspThrLys

                          225
                          |
GAAGGCATCCTGCAGTATTGCCAAGAAgtaagtcctgtccggtggctagcaattcacgtt
GluGlyIleLeuGlnTyrCysGlnGlu
                        75

ggatcacatgcatttgttttcaaaaatttaacttctgttattttgcatcagtattttaa

cccctacagtaaaaatcttggttcctaatgattcaccataccattaatatatttatttgca

ttaccctatgatatacatataaatgtttttaaaattatgatgtcgtattatgaccatcac

taaacagtagtttaagatgtcacagcactttttttatttcctcactctgtcacccaggttg

gagtgcagtggcaagattatggcttactgtagccttgacctactgggctcaagcaatcct

cccacctcagcctcccaagttcctggactataggcacatgccaccaggcctgactgattt

tttttaatttttagtagagatggggtctccttgtgttggccaggctgtctgaaactcctg

                                             226
                                             |
ggctcaagcgatcctc...................tgctctcccagGTCTACCCTGAACTG
                                             ValTyrProGluLeu
                                             76

CAGATCACCAATGTGGTAGAAGCCAACCAACCAGTGACCATCCAGAACTGGTGCAAGCGG
GlnIleThrAsnValValGluAlaAsnGlnProValThrIleGlnAsnTrpCysLysArg

                                                       355
                                                       |
GGCCGCAAGCAGTGCAAGACCCATCCCCACTTTGTGATTCCCTACCGCTGCTTAGgtgag
GlyArgLysGlnCysLysThrHisProHisPheValIleProTyrArgCysLeuV
                                                       119

ccg.......................cttgaagtctatctttccttgatgtcttctgcg

                                                      356
                                                      |
gtaagaacactgtgatacagatggaatgacgggaagtggtttcctttctttcagTTGGT
                                                      alGly
                                                        120

GAGTTTGTAAGTGATGCCCTTCTCGTTCCTGACAAGTGCAAATTCTTACACCAGGAGAGG
GluPheValSerAspAlaLeuLeuValProAspLysCysLysPheLeuHisGlnGluArg

                                               468
                                               |
ATGGATGTTTGCGAAACTCATCTTCACTGGCACACCGTCGCCAAAGAGgtaccagcccat
MetAspValCysGluThrHisLeuHisTrpHisThrValAlaLysGlu
                                             156

aaattctttcttattgcaaagtgaagatttcctggggacgtgctt...............

.....ttgtgaaattggttcctaatatattgggtctgcatgttgattattttatgtggag

ttttcttacaatgaaacacatctactctaccactcactgtttttctccttacactttgtag

469
|
ACATGCAGTGAGAAGAGTACCAACTTGCATGACTACGGCATGTTGCTGCCCTGCGGAATT
ThrCysSerGluLysSerThrAsnLeuHisAspTyrGlyMetLeuLeuProCysGlyIle
157

GACAAGTTCCGAGGGGTAGAGTTTGTGTGTTGCCCACTGGCTGAAGAAAGTGACAATGTG
AspLysPheArgGlyValGluPheValCysCysProLeuAlaGluGluSerAspAsnVal

GATTCTGCTGATGCGGAGGAGGATGACTCGGATGTCTGGTGGGGCGGAGCAGACACAGAC
AspSerAlaAspAlaGluGluAspAspSerAspValTrpTrpGlyGlyAlaAspThrAsp
```

```
             662
             |
TATGCAGATGGAAGgtaaggtggcctttgtgttcagcctcagagatgctgaaacatcttg
TyrAlaAspGlySe
             221

                                                   663
                                                   |
tatggagtatttgtatcctgtaaattaatc.....................TGAAGACAAA
                                                   rGluAspLys
                                                    222

GTAGTAGAAGTAGCAGAGGAGGAAGAAGTGGCTGAGGTGGAAGAAGAAGAAGCCGATGAT
ValValGluValAlaGluGluGluGluValAlaGluValGluGluGluGluAlaAspAsp

GACGAGGACGATGAGGATGGTGATGAGGTAGAGGAAGAGGCTGAGGAACCCTACGAAGAA
AspGluAspAspGluAspGlyAspGluValGluGluGluAlaGluGluProTyrGluGlu

GCCACAGAGAGAACCACCAGCATTGCCACCACCACCACCACCACCACAGAGTCTGTGGAA
AlaThrGluArgThrThrSerIleAlaThrThrThrThrThrThrThrGluSerValGlu

            865 *
            |
GAGGTGGTTCGAG.........................aaatacggctttctattaaa
GluValValArgV
            289

cgagtggattattctgttgttgttggctttttttctcaaacctccttctcttctacttt

  * 866
    |
atagTTCCTACAACAGCAGCCAGTACCCCTGATGCCGTTGACAAGTATCTCGAGACACCT
    alProThrThrAlaAlaSerThrProAspAlaValAspLysTyrLeuGluThrPro
      290

GGGGATGAGAATGAACATGCCCATTTCCAGAAAGCCAAAGAGAGGCTTGAGGCCAAGCAC
GlyAspGluAsnGluHisAlaHisPheGlnLysAlaLysGluArgLeuGluAlaLysHis

                 999
                 |
CGAGAGAGAATGTCCCAGgtaagtctggctcttccatcatt...................
ArgGluArgMetSerGln
               333

............aaggatcgaaactgatgaactttaaattcaaatgtttccttaattta

   1000
   |
tagGTCATGAGAGAATGGGAAGAGGCAGAACGTCAAGCAAAGAACTTGCCTAAAGCTGAT
   ValMetArgGluTrpGluGluAlaGluArgGlnAlaLysAsnLeuProLysAlaAsp
   334

                 1074
                 |
AAGAAGGCAGTTATCCAGgtaaaacctgaacccatttccta.................aaa
LysLysAlaValIleGln
               358

acgtagaattagtcaatgttggaatgactatgcagtctttaggatactttttagcactag

aagaaaatggaataggatgtctttttttagaagacttgaaattgctgcttcatcctactta

                                    1075
                                    |
ttcagtccccatggacatatgtgtctatgatggcagCATTTCCAGGAGAAAGTGGAATCT
                                    HisPheGlnGluLysValGluSer
                                    359

TTGGAACAGGAAGCAGCCAACGAGAGACAGCAGCTGGTGGAGACACACATGGCCAGAGTG
LeuGluGlnGluAlaAlaAsnGluArgGlnGlnLeuValGluThrHisMetAlaArgVal

GAAGCCATGCTCAATGACCGCCGCCGCCTGGCCCTGGAGAACTACATCACCGCTCTGCAG
GluAlaMetLeuAsnAspArgArgArgLeuAlaLeuGluAsnTyrIleThrAlaLeuGln

              1233
              |
GCTGTTCCTCCTCGGgtaggtctcgctgcagccgagttcacacttcaggtcacagcacag
AlaValProProArg
             411

acagtaagggtggggcactgggaactggaagccatacaaaagaatgagggagaaatgcct

tgagcactgttattcagaggttcaacccctgtccattccatcttgaaggtcaaagggtca

                                                   1234
                                                   |
caggggcagctacctccacaaggtcatctctacacagcagg..........CCTCGTCAC
                                                   ProArgHis
                                                   412

GTGTTCAATATGCTAAAGAAGTATGTCCGCGCAGAACAGAAGGACAGACAGCACACCCTA
ValPheAsnMetLeuLysLysTyrValArgAlaGluGlnLysAspArgGlnHisThrLeu

                                                         1362
                                                           |
AAGCATTTCGAGCATGTGCGCATGGTGGATCCCAAGAAAGCCGCTCAGATCCGGTCCCAG
LysHisPheGluHisValArgMetValAspProLysLysAlaAlaGlnIleArgSerGln
                                                         454

                            1363
                            |
.....................tgatgcagGTTATGACACACCTCCGTGTGATTTATGAGCGC
                            ValMetThrHisLeuArgValIleTyrGluArg
                            455
```

```
ATGAATCAGTCTCTCTCCCTGCTCTACAACGTGCCTGCAGTGGCCGAGGAGATTCAGGAT
MetAsnGlnSerLeuSerLeuLeuTyrAsnValProAlaValAlaGluGluIleGlnAsp

     1462
      |
GAAGTTGgtaagtaagcttttcttttatct............................tt
GluValA
      488

gaatgccatgtgcctcagttttcacatctgtaaaagggagatgataatggtacctatgtc

atggctctaaacgcgatcatgcacgtgaaagcagttgaagtcttgcctggcagaagtaaa

tggtggctgctgctgctgctgctgttgtgattgttgttactcaccaaagagatggttttg

       1463
          |
tttggtttagATGAGCTGCTTCAGAAAGAGCAAAACTATTCAGATGACGTCTTGGCCAAC
          spGluLeuLeuGlnLysGluGlnAsnTyrSerAspAspValLeuAlaAsn
            489

ATGATTAGTGAACCAAGGATCAGTTACGGAAACGATGCTCTCATGCCATCTTTGACCGAA
MetIleSerGluProArgIleSerTyrGlyAsnAspAlaLeuMetProSerLeuThrGlu

ACGAAAACCACCGTGGAGCTCCTTCCCGTGAATGGAGAGTTCAGCCTGGACGATCTCCAG
ThrLysThrThrValGluLeuLeuProValAsnGlyGluPheSerLeuAspAspLeuGln

                                                   1684
                                                   |
CCGTGGCATTCTTTTGGGGCTGACTCTGTGCCAGCCAACACAGAAAACGAAGgtaagagt
ProTrpHisSerPheGlyAlaAspSerValProAlaAsnThrGluAsnGluV
                                                   562

                                        1685
                                        |
cccctgagccagcaagg.......................TTGAGCCTGTTGATGCCCGC
                                        alGluProValAspAlaArg
                                          563

                                  1738
                                  |
CCTGCTGCCGACCGAGGACTGACCACTCGACCAG..........................
ProAlaAlaAspArgGlyLeuThrThrArgProG
                                  580

tttaagagtcatttatcaattttcttctaacttcaggcctagaaagaagttttgggtagg

ctttgtcttacagtgttattatttatgagtaaaactaattggttgtcctgcatactttaa

               1739
                |
ttatgatgtaatacagGTTCTGGGTTGACAAATATCAAGACGGAGGAGATCTCTGAAGTG
                lySerGlyLeuThrAsnIleLysThrGluGluIleSerGluVal
                  581
```

```
                                                         1839
                                                         |
AAGATGGATGCAGAATTCCGACATGACTCAGGATATGAAGTTCATCATCAAAAATTGgta
LysMetAspAlaGluPheArgHisAspSerGlyTyrGluValHisHisGlnLysLeu
                                                       613

cgtaaaataatttacctctttccactactgtttgt.......................aaag

taagcatcaaatatttgaccaaccagttgggcagagaatatactgaaactttttatataa

cctcatccaaatgtccccgtcatttaagaaatgaaattcttctaattgcgtttataaatt

                                                      1840
                                                      |
gtaaattatattgcatttagaaattaaattctttttcttaatttgttttcagGTGTTC
                                                    ValPhe
                                                    614

TTTGCAGAAGATGTGGGTTCAAACAAAGGTGCAATCATTGGACTCATGGTGGGCGGTGTT
PheAlaGluAspValGlySerAsnLysGlyAlaIleIleGlyLeuMetValGlyGlyVal

GTCATAGCGACAGTGATCGTCATCACCTTGGTGATGCTGAAGAAGAAACAGTACACATCC
ValIleAlaThrValIleValIleThrLeuValMetLeuLysLysLysGlnTyrThrSer

                    1986
                    |
ATTCATCATGGTGTGGTGGAGgtaggtaaacttgactgcatgtttccaagtgggaattaa
IleHisHisGlyValValGlu
                  662

gactatgagagaattaggcttagcttttgctaagaactagctaagtatctcttttaaaa

aacgaatcagtgtgcttccatgatgcttgggttacagttgttctttctt..........

.tgggcaaaggcgttctgctccaagatgtcaaagtggggagaaaattatgggtgttctg

caatcttggcaaaagaagagtactgtactccttatctttttactgcttctccatgttcac

                                                   1987
                                                   |
ccttaaaagattgatttttatttttactcagctctcctcttgtttttcagGTTGACGCC
                                                   ValAspAla
                                                   663

GCTGTCACCCCAGAGGAGCGCCACCTGTCCAAGATGCAGCAGAACGGCTACGAAAATCCA
AlaValThrProGluGluArgHisLeuSerLysMetGlnGlnAsnGlyTyrGluAsnPro

                                2088
                                |
ACCTACAAGTTCTTTGAGCAGATGCAGAACTAG
ThrTyrLysPhePheGluGlnMetGlnAsnEnd
                              695
```

Figure 1: Exon−intron boundaries of the human PAD-gene.
The DNA sequence of the 16 PreA4$_{695}$ exons are written in capital letters and numbered as the corresponding PreA4$_{695}$-cDNA (8). The DNA sequences of the partially sequenced PreA4$_{695}$ introns are written in small letters and the points interrupting the intronal sequences indicate regions within the introns which have not been sequenced. The varying number of points within the different introns do not correspond to the size of these gaps. The asterisk's at the bases 865 and 866 indicate the exon−intron boundary where the further exons mentioned in the text may be inserted. The amino acid sequence written under the exons corresponds to the deduced amino acid sequence of the PreA4$_{695}$-cDNA.

chromosome 21 library (*Hind*III and *Eco*RI fragments in *lambda* charon 21A) from the American Type Culture Collection (ATCC) in Rockville, Maryland, USA, c) a human genomic library (*lambda* vector L47.1, courtesy of Dr B. Horsthemke, University of Essen), and d) a human leukocyte genomic library (*lambda*-EMBL 3) purchased from Genofit, Heidelberg. Different DNA fragments were isolated from the PreA4$_{695}$cDNA (8) and used as screening probes. Hybridization conditions for the first three libraries mentioned above have been described (7). Hybridizations with the human genomic library from Genofit were performed in 6×SSC, 5×Denhardt's solution, 0.5% SDS, 0.05% Na-pyrophosphate, 100μg/ml salmon sperm DNA at 65° with 5 × 10^5c.p.m/ml of randomly primed probe (13). Putative positive *lambda*-clones were rescreened under the same conditions and the *lambda*-DNA was isolated as described (14).

Treatment of positive lambda clones

In each case the DNA inserts of positive *lambda* clones from the chromosome 21 libraries were cloned completely into the plasmid vector pUC19 (15). The DNA inserts of positive *lambda* clones from the human genomic libraries were digested with different restriction endonucleases and subcloned into pUC19. Exon-containing fragments were detected by dideoxy sequencing (16,17) using exon-specific primers.

Polymerase chain reaction (PCR)

Amplifications with *Taq*I-polymerase (18) were performed in 100µl reaction mixtures containing 1µg genomic DNA (human embryonic liver) in 50mM KCl, 10mM Tris-HCl, pH 8.5, 2.5mM $MgCl_2$, each primer at 200nM, each dNTP (dATP, dCTP, dGTP, dTTP) at 200µM, gelatine at 200µg/ml and 2.5 units of polymerase (Cetus, Perkin-Elmer). The samples were overlaid with several drops of mineral oil and subjected to 40 cycles of amplification as follows. The samples were heated from 70° to 95° over a 2-minute period (denaturation), cooled to 55° over 2 minutes (annealing), heated to 70° for 1 minute (extension reaction). Thermal cycling was performed in a programmable heatblock (Perkin-Elmer, Cetus Instruments). After the final extension step, the samples were precipitated with ethanol and resuspended in 100µl TE buffer (19). 80µl of each sample was resolved on a 1.2% seaplaque-agarose gel and the fragment of interest was isolated as described (19). The amounts of synthesized fragments were approximately 200–500ng in each case. 50ng of each sample was used for cloning into the plasmid vector pUC19. The cloning of the exon-representing products was confirmed by sequence analysis.

DNA-sequencing

The sequences presented in figure 1 were derived by the chain termination method (16) using Klenow polymerase on single-stranded denatured plasmid DNA templates (17). Exon-specific synthetic primer-oligomers were synthesized on a Model 380 A DNA synthesizer (Applied Biosystems).

RESULTS AND DISCUSSION

Exon-specific clones isolated from the genomic libraries

The following fragments were isolated from the chromosome 21 libraries and cloned completely into the plasmid vector pUC19 (15): H1.30(2.8kb,exon1), E6.BA(1.8kb,exon2), H3.31(4.5kb,exon3), H2.31(7.5kb,exons4,5), 4A(7.0kb,exon7), H1.41(3.8kb,exons 8,9), APC1(1.2kb, exon 14), E4.5(2.8kb, exon 15), H1.23(7.0kb, exon 16). The lengths of the different fragments and the numbers of the corresponding exons are shown in brackets. Clone P1.21(1.2kb,exon12) was isolated from the human genomic library from B.Horsthemke and represents only a part of a larger insert from the corresponding *lambda* clone. Clone HG440(appr.15kb) was isolated from the human leukocyte genomic library (Genofit) and a 4.5kb fragment containing exon 11 was subcloned in pUC19.

Exon-specific clones synthesized by PCR (18)

Lambda clones specific for the exons 6,10 and 13 could not be detected in the genomic libraries mentioned above. Three *lambda* clones specific for exon 6 were isolated from the human leukocyte genomic library (Genofit) but up to now they have not been subcloned into pUC19. In order to find out whether the putative exons 6,10 and 13 were interrupted by further introns we probed the DNA by the PCR amplification method (18) as follows. The exon–intron boundaries of exons 5 and 7,9 and 11, 12 and 14, respectively had been determined by the analysis of neighbouring genomic clones. So we used two synthetic oligonucleotides with their 5′ ends corresponding to the first or last position of the putative exon. The products of amplification procedures were shown to have the same size as the corresponding cDNA fragments on a 2% agarose gel (data not shown), and sequence analysis of the cloned fragments revealed no further introns.

Exon – intron boundaries determined by DNA sequence analysis

Partial sequence analysis of the various genomic clones revealed that the PreA4_{695} transcript is a splicing product containing 16 exons (figure 1). In each case the exon – intron boundary as well as the whole exon was sequenced. As expected (20), each intron starts with a 'GT' and ends with a 'AG'.

Only two of the genomic clones isolated from the genomic libraries contain two exons (clones H2.31, 7.5Kb, exons 4 and 5, and H1.41, 3,8kb,exons 8 and 9). All the other clones contain one exon each, even HG440 (appr. 15 Kb) and P1.21 (appr. 14 Kb). Hybridization experiments (data not shown) showed that the neighbouring genomic clones did not overlap, thus the size of the gene cannot yet be determined. From the known insert sizes a minimal length of 50 Kb can be calculated.

Recently, three groups reported the cloning of preA4-cDNAs containing an extra exon encoding a protease-inhibitor like sequence (9,10,11). Kitaguchi et al. (10) isolated a cDNA containing the exon coding for the protease-inhibitor like sequence together with a 3'-adjacent small exon coding for a peptide similar in sequence to the MRC OX2 antigen (21). They proposed (10) that the three different preA4-mRNA's are due to alternative splicing of a single PAD gene. For this form the PAD gene provides 18 exons. The use of the PreA4_{695} cDNA in the identification of the genomic structure of the PAD gene would have failed to reveal the trypsine inhibitor coding exon or any further exons. In figure 1 the positions of the exon – intron boundaries where the trypsin-inhibitor-like exon and the other additional exon (10) have been found are marked by asterisk's (bases 865, 866).

Exon – intron boundaries and the structure of the PreA4_{695} protein.

Comparison of the exon – intron-structure of the PAD-gene to the deduced protein sequence shows that exon 1 contains the coding region for the signal peptide (22). Mita et al. (23) recently reported a cDNA coding for a further 73 amino acids at the N-terminus of the precursor. The significance of this finding is still unclear. Exons 2,3,4 and 5 span the cysteine-rich region, and exons 5 and 6 provide the highly negatively charged domain (8). The amino acid sequence encoded by exon 6 shows some similarity (31 %) to human prothymosin alpha-1 (24) but this similarity is mainly based on the acidic amino acids in these two peptides. The protein sequence encoded by exons 7 to 13 contains the two putative N-glycosylation sites (8,25) of the PreA4_{695}-protein. The amyloid A4 protein extends across the border between exons 14 and 15, a fact that supports the idea that accummulation of this peptide in the brain tissue is due to degradation of the precursor and not to aberrant splicing. Finally the putative transmembrane region is completely contained in exon 15 and most of the putative cytoplasmic domain is coded for by exon 16.

ACKNOWLEDGEMENTS

We thank A.Starzinski-Powitz for the human embryonic DNA, B.Horsthemke for the human genomic library, J.Hesselbach and A.Lee for excellent technical assistance, and B.Kisters for help with the computers. This work was supported by a grant of the Fritz Thyssen Stiftung to B.M.H. and grants of the Deutsche Forschungsgemeinschaft through SFB 317, the Bundesministerium für Forschung und Technologie and the Fonds der chemischen Industrie to K.B.

REFERENCES

1. Alzheimer, A. (1907), Allg.Z.Psychiat.*64*, 146 – 148
2. Katzman,R. (1986), N. Engl. J. Med. *314*, 964 – 973.

3. Masters, C.L., Simms,G., Weinmann,N.A., Multhaup,G., McDonald,B.L. and Beyreuther,K. (1985), Proc.Natl.Acad.Sci.USA *82*, 4245–4249
4. Masters,C.L., Multhaup,G., Simms, G., Pottgiesser,J., Martins,R.N. and Beyreuther,K. (1985), EMBO J.*4*, 2757–2763
5. Katzman,R.(ed.) (1983), Biological Aspects of Alzheimer's Disease (Banbury Report 15), Cold Spring Harbor Laboratory, New York
6. Glenner,G.G. and Wong,C.W. (1984), Biochem. Biophys. Res. Commun. *120*, 1131–1135
7. Salbaum,J.M., Weidemann,A., Lemaire,H.G., Masters,C.L. and Beyreuther,K. (1988), EMBO J. *7*, 2807–2813
8. Kang,J., Lemaire,H.G., Unterbeck,A., Salbaum,J.M., Masters,C.L., Grzeschik,K.H., Multhaup,G., Beyreuther,K. and Müller-Hill,B. (1987), Nature *325*, 733–736
9. Tanzi,R.E., McClatchey,A.I., Lamperti,E.D., Villa-Komaroff,L., Gusella,J.F. and Neve,R.L. (1988), Nature *331*, 528–530
10. Kitaguchi, N., Takahashi, Y., Tokushima,Y., Shiojiri,S., and Ito,H. (1988), Nature *331*, 530–532
11. Ponte,P., Gonzalez-DeWhitt,P., Schilling,J., Miller,J., Hsu,D., Greenberg,B., Davis,K., Wallace,W., Lieberburg,I., Fuller,F. and Cordell,B. (1988), Nature *331*, 525–527
12. Schubert,D., Schroeder,R., LaCorbiere,M., Saitoh,T., and Cole,G. (1988), Science *241*, 223–226.
13. Feinberg,A.P. and Vogelstein,B. (1984), Anal.Biochem. *137*, 266
14. Lambda Sorb^{TM}Phage Adsorbent Kit, Promega Biotec, Madison, Wi 53711, USA.
15. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985), Gene *33*, 103–119
16. Sanger,F., Nicklen,S. and Coulson,A.R. (1977), Proc.Natl.Acad.Sci. USA *74*, 5463–5467
17. Chen,E.J. and Seeburg,P.H. (1985), DNA *4*, 165–170
18. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988), Science *239*, 487–491
19. Maniatis,T., Fritsch,E.F., Sambrook,J., Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY) (1982).
20. Leder,A., Miller,H.J., Hamer,D.H., Seidman,J.G., Norman,B., Sullivan,M. and Leder,P. (1978), Proc.Natl.Acad.Sci. USA *75*, 6187–6191
21. Clark,M.J., Gagnon,J., Williams,A.F. and Barclay,A.N. (1985), EMBO J. *4*, 113–118
22. Blobel,G. and Dobberstein,B. (1975), J.Cell,Biol.*67*, 852–862
23. Mita,S., Sadlock,J., Herbert,J., and Schon,E.A. (1988), Nucleic Acids Res. *16*, 9351.
24. Goodall,G.J., Dominguez,F. and Horecker,B.L. (1986), Proc.Natl. Acad.Sci.USA *83*, 8926–8928
25. Dyrks,T., Weidemann,A., Multhaup,G., Salbaum,J.M., Lemaire,H.G., Kang,J., Müller-Hill,B., Masters,C.L. and Beyreuther,K. (1988), EMBO J. *7*, 949–957

Exhibit C

A new A4 amyloid mRNA contains a domain homologous to serine proteinase inhibitors

P. Ponte*, P. Gonzalez-DeWhitt*, J. Schilling*, J. Miller*, D. Hsu*, B. Greenberg*, K. Davis†, W. Wallace†, I. Lieberburg†, F. Fuller* & B. Cordell*

* California Biotechnology, Inc., 2450 Bayshore Parkway, Mountain View, California 94043, USA
† Department of Psychiatry, Arthur M. Fishberg Research Center in Neurobiology, Mount Sinai School of Medicine, 1 Gustave Levy Place, New York, New York 10029, USA

The amyloid proteins isolated from neuritic plaques and the cerebrovasculature of Alzheimer's disease are self-aggregating moieties termed A4 protein[1] and β-protein[2,3], respectively. A putative A4 amyloid precursor (herein termed $A4_{695}$) has been characterized by analysis of a human brain complementary DNA[4]. We report here the sequence of a closely related amyloid cDNA, $A4_{751}$, distinguished from $A4_{695}$ by the presence of a 168 base-pair (bp) sequence which adds 57 amino acids to, and removes one residue from, the predicted $A4_{695}$ protein. The peptide predicted from this insert is very similar to the Kunitz family of serine proteinase inhibitors. The two A4-specific messenger RNAs are differentially expressed: in a limited survey, $A4_{751}$ mRNA appears to be ubiquitous, whereas $A4_{695}$ mRNA has a restricted pattern of expression which includes cells from neuronal tissue. These data may have significant implications for understanding amyloid deposition in Alzheimer's disease.

Two full-length A4-specific cDNA clones were isolated from a λgt10 library and were characterized (Fig. 2). Both cDNA clone sequences were identical to the $A4_{695}$ sequence[4] except for a 168 bp insert (Fig. 2) resulting in an insertion of 57 amino acids and the removal of one residue of the predicted 695 amino-acid sequence of the $A4_{695}$ protein[4]. This 57 amino-acid sequence was compared to the protein sequence database of the Protein Identification Resource. The search revealed extensive similarity between the insert and the Kunitz family of low relative molecular mass serine proteinase inhibitors (Fig. 1): for the proteins compared in the figure the identity ranges from 33 to 48%. This similarity suggests that the $A4_{751}$ insert could act as a proteinase inhibitor.

To investigate the expression of $A4_{751}$ and $A4_{695}$ mRNAs, two synthetic oligonucleotides were synthesized. One oligonucleotide corresponds to 60 bases of the 168-bp insert (insert probe) and the other oligonucleotide spans the insert point, possessing 15 bases on either side (junction probe). Under hybridization wash conditions favouring longer oligonucleotides the junction probe is stable when bound to $A4_{695}$ mRNA (30 bp perfect match), but unstable when bound to $A4_{751}$ mRNA (15 bp match). These probes were used to characterize the $A4_{695}$ and $A4_{751}$ cDNA representation in four cDNA libraries prepared from human RNA derived from a SV40-transformed fibroblast cell line, lymphocytes, and normal and Alzheimer's disease brain (see legend to Fig. 2). Insert-positive clones were identified in all libraries; however, only the brain libraries contained junction-positive clones. These results suggested that the expression of the A4-specific transcripts might be differentially regulated. We examined this possibility by characterizing steady-state mRNA levels in several different sources.

Polyadenylated RNA was prepared from human cultured cell lines and from human brain sections. The cultured cells used were HeLa, MRC5 and IMR-32 (see legend to Fig. 3). Tissue samples represented normal cerebellum, frontal and parietal cortex and frontal cortex from an Alzheimer's disease patient. An identical pair of Northern blots were prepared and probed with either the insert or junction oligonucleotide. Both blots were then stripped and reprobed with an actin cDNA probe as an internal control. The results of the Northern analysis (Fig. 3) reveal that each oligonucleotide specifically detects a 3.2-3.4 kilobase (kb) mRNA, consistent with the size of the mRNA encoding the putative A4 amyloid precursor[4-7]. The junction probe, however, hybridizes to only a subset of the samples, those of neural origin (Fig. 3*a*, lanes 1, 6-9). No mRNA is detected in the HeLa or MRC5 samples using the junction probe even in the deliberately overexposed autoradiograph shown. In contrast, the insert probe detects sequences in all samples (Fig. 3*b*). The relative expression levels of the two mRNAs were also examined by RNA dot blot using total RNA prepared from several additional cell lines (P.G.-D. and P.P., data not shown). A human glioblastoma line (U-87-MG) and two human neuroblastoma lines (SK-N-MC and SK-N-SH) all contain mRNA species which hybridize to the insert probe, but none of these samples show significant hybridization to the junction probe above background levels (using MRC5 RNA as the negative control).

At least two types of A4-specific mRNA (and presumably protein) exist which are differentially regulated. We have also determined that the inserted sequence is entirely and exclusively encoded on a separate exon (J.M., F.F. & D.H. unpublished observation). From information on Alzheimer's disease amyloid

Protein	290 300 310 320 330 340	% Identity	Ref.
$A4_{751}$ insert	EVCSEQAETGPCRAMISRWYFDVTEGKCAPFFYGGCGGNRNNFDTEEYCMAVCGSAI	-	-
Bovine pancreatic inhibition	DFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGAI	47	30
Bovine serum inhibition	DFCLEPPYTGPCKAAMIRYFYNAKAGFCETFVYGGCRAKSNNFKSAEDCMRTCGGA	42	31
Bovine colostrum inhibition	DLCQLPQARGPCKAALLRYFYDSTSNACEPFTYGGCQGNNDNFETTEMCLRICEPPQ	42	32
Human ITI Domain I	DSCQLGYSAGPCMGMTSRYFYNGTSMACETFQYGGCMGNGNNFVTEKECLQTCRTV	40	33
Domain II	AACNLPVIRGPCRAFIQLWAFDAVKGKCVLFPYGGCQGNGNKFYSEKECREYCGVPG	46	
Bovine ITI Domain I	DSCQLDYSQGPCLGLFKRYFYNGTSMACETFLYGGCMGNLNNFLSQKECLQTCRTV	33	34
Domain II	EACNLPIVQGPCRAFIQLWAFDAVKGKCVRFSYGGCKGNGNKFYSQKECKEYCGIPG	46	

Fig. 1 Comparison of the amino-acid sequences of the $A4_{751}$ insert with Kunitz proteinase inhibitors. The original computer search was conducted against the Protein Identification Resource database using the FASTP routine of Pearson and Lipman[22]. Only mammalian proteins are compared although homologous proteins which are members of this family have also been isolated from reptiles, molluscs and coelenterates (reviewed in ref. 23). ITI, inter-α trypsin inhibitor. Only the 57 residues that overlap with the $A4_{751}$ insert are displayed. Numbering is based on the $A4_{751}$ predicted protein sequence as defined in Fig. 1. The 13 residues boxed are invariant in the other members of this inhibitor family and are all conserved in the $A4_{751}$ insert. The percentage identity of each sequence with the $A4_{751}$ insert sequence is shown. Human and bovine ITI domain I is glycosylated at Asn-310.

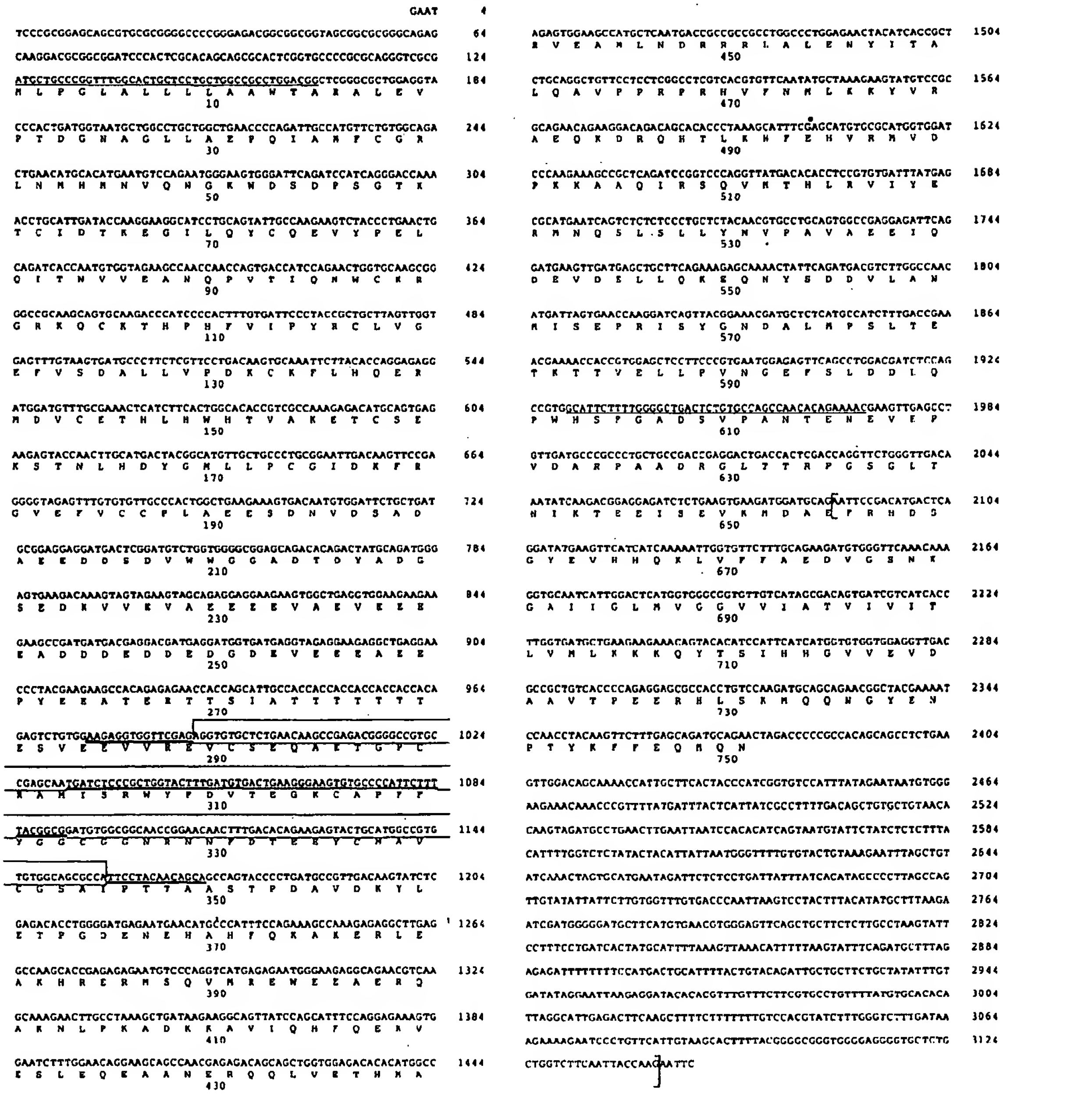

Fig. 2 Nucleotide sequence and deduced 751 amino-acid sequence of the $A4_{751}$ cDNA. Numbering of nucleotides begins with the first base of the clone and proceeds in a 5' to 3' direction for a total of 3,148 bases. The inserted 167 bp domain is boxed (nucleotides 990-1,157); the insert-specific oligonucleotide (nucleotides 1,032-1,091) and deletion-specific oligonucleotide (nucleotides 975-989 plus 1,158-1,172) are underlined in bold; oligonucleotides used in cDNA isolation are underlined (nucleotides 125-167 and 1,930-1,972); a partial A4-specific cDNA clone (nucleotides 2,088-3,143) previously isolated from a normal brain library (unpublished data) is bracketed.

Methods. Similarity studies and formatting were carried out using computer programs from Intelligenetics and the Wisconsin Genetics Computer Group, respectively. Complementary DNAs were isolated from a library constructed using RNA prepared from a SV40-transformed human fibroblast line described by Wolf and Rotter[17]. Oligo(dT)-primed double-stranded cDNA was synthesized from total poly(A)$^+$ RNA, then ligated with linkers and cloned into the *Eco*R1 site of the λgt10 vector; the brain and lymphocyte libraries were obtained from Clonetech, Palo Alto, CA. About 1×10^6 phage from the SV40-transformed fibroblast library were screened by hybridization[18] with synthetic oligonucleotides labelled by kinasing[19] and with a cDNA probe labelled by nick-translation[20]. Twelve hybridization-positive (for all probes) phage clones were characterized for insert length from which two were subcloned into M13 vectors for sequence analysis by the dideoxy chain termination procedure[21]. Both strands of each cDNA were sequenced using synthetic primers.

gene copy number[8-10] we infer that the two RNA species arise by alternative splicing of a single transcriptional unit. Two A4-specific transcripts of 3.2 and 3.4 kb have been previously identified in normal adult human cortex and in the cortex of Alzheimer's disease patients[4]. The presence or absence of the 168-bp insert, or the use of two existing polyadenylation sites in the 3'-untranslated region[4] may explain the differences in message length. Examples of brain-specific alternative RNA splicing have previously been reported[11-13].

It is premature to conclude that $A4_{695}$ mRNA is expressed only in cells of neuronal origin. In fact, preliminary experiments indicate that HL-60, a promyelocytic leukaemia line, expresses both forms (P.P., unpublished observation). *In situ* hybridization analysis will be necessary to obtain a complete understand-

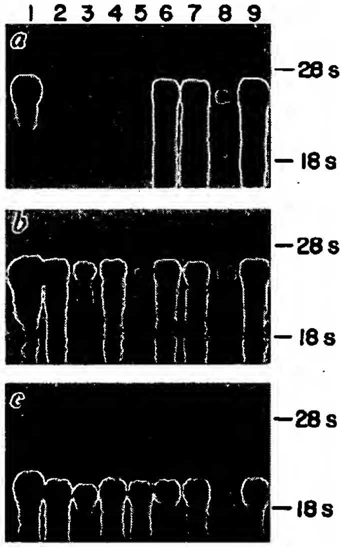

Fig. 3 Northern analysis of $A4_{695}$ and $A4_{751}$ mRNA species. Total or cytoplasmic RNA was isolated from human adult brain sections or cells propagated in culture. The cultured cells are HeLa, an epithelial-like cell derived from a cervical carcinoma; MRC5, a primary fibroblast line from fetal lung tissue; IMR-32, a mixed culture developed from a peripheral neuroblastoma with the predominant cell of neuroblast-like characteristics and the minor type resembling a hyaline fibroblast. All cell lines used are available from the American Type Culture Collection. Total RNA preparations from brains were isolated by the guanidine thiocyanate/LiCl method of Cathala *et al.*[24]. A cytoplasmic fraction from confluent monolayers of HeLa and MRC5 cells was prepared by lysis of the cells in a hypotonic buffer of 10 mM Tris *p*H 7.5, 10 mM KCl, 1 mM $MgCl_2$, 1% Triton-X100, 0.1% sodium deoxycholate for 10 min at 4 °C, homogenization in a Dounce with a 'B' pestle for 10 strokes, and pelleting the nuclei at 2,500*g* for 5 min. The supernatant was mixed with guanidine thiocyanate, and RNA was pelleted through CsCl by the method of Chirgwin *et al.*[25]. Total RNA from the same cells was obtained by lysis of the cells directly into guanidine thiocyanate followed by CsCl fractionation as above. RNA (100–400 μg) was fractionated by oligo(dT) cellulose chromatography by the method of Aviv and Leder[26]. Samples containing $poly(A)^+$ RNA were divided into two aliquots, run on duplicate formaldehyde/1% agarose gels, and blot-transferred to nitrocellulose by the method of Thomas[27]. Nitrocellulose filters were hybridized either to the junction oligonucleotide (5'-CTGCTGTTGTAGGAACTCGAACCACCTCTT-3') or the insert oligonucleotide

(5'-CGCCGTAAAAGAATGGGGCACACTTCCCTTCAGTCA
CATCAAAGTACCAGCGGGAGATCA-3')

which had been labelled with a 15–30 nucleotide-long homopolymeric tail using [α-^{32}P]dCTP and terminal deoxynucleotidyl transferase as previously described[28]. The buffers used for hybridization contained 10% dextran sulphate, 5 × Denhardt's reagent, 50 mM sodium phosphate (*p*H 6.7), with SSC and formamide concentrations dependent on the probe used—junction oligonucleotide: 20% formamide, 6 × SSC; insert oligonucleotide: 34% formamide, 4 × SSC; actin: 50% formamide, 4 × SSC. Blots were washed to a final stringency of 1 × SSC at 55 °C, and exposed to Kodak X-AR film for 3 days. Both blots were then stripped and rehybridized to a human β-actin cDNA insert[29] which had been labelled by nick-translation with [α-^{32}P]dCTP[20]. Re-exposure (6 h) of the filter used in *a* is shown in *c*; re-exposure of filter shown in *b* is very similar and is not presented. The RNA samples displayed are: Lane 1, total IMR-32; 2, total MRC5; 3, total HeLa; 4, cytoplasmic MRC5; 5, cytoplasmic HeLa; 6, normal cerebellum; 7, normal frontal cortex; 8, AD frontal cortex; 9, normal parietal cortex. *a*, RNAs hybridized with the junction oligonucleotide; *b*, RNAs hybridized with the insert oligonucleotide; and *c*, RNAs hybridized with human β-actin cDNA. Ribosomal RNAs used as internal size markers are shown. The normal brain samples were obtained from two different individuals, 98 years of age (frontal cortex and cerebellum samples) and 60 years of age (parietal cortex sample), both lacking clinical signs of dementia. The Alzheimer's disease sample was obtained from a 97-year-old patient displaying overt clinical indications of the disease. The tissue samples were obtained 4, 14 and 5 h post-mortem from the 98, 60 and 97-year-old individuals, respectively. Because different amounts of RNA may be contained in each sample, determination of the relative amounts of each A4-specific species in the different samples is not possible.

ing of the distribution and expression levels of these two A4-specific mRNAs in normal and in pathological conditions such as Alzheimer's disease. The single Alzheimer's disease cortex sample examined here shows no striking differences in the relative amounts of each mRNA compared to the normal cortex samples.

The similarity of the 57-amino-acid insert to a family of proteinase inhibitors suggests that it may have a similar function. The biosynthesis and physiological role of this proteinase inhibitor family are not well understood. Bovine pancreatic trypsin inhibitor is derived from a larger precursor[14] and the other members of the family may be as well. In each case, the functional inhibitor is ~57 amino acids long or a dimer of two such units. The precursor of the human inter-α-trypsin inhibitor may be functionally active[15]. Similarly, the entire $A4_{751}$ protein might function as a proteinase inhibitor or the putative inhibitor domain may be proteolytically excised and function independently. The physiological role of other serine proteinase inhibitors is the regulation of a single proteinase[16]. The analysis of the physiological function of the $A4_{751}$ insert and the identification of its *in vivo* target must await the production of the pure protein.

We thank D. Wolf for the SV40-transformed fibroblast library, E. Stoelting for photographic work, C. Boyle and L. Brown for manuscript preparation, C. Bryant and T. Scheuermann for technical assistance and J. Feldman and L. Berg for their efforts in the computer facility. This work was supported by California Biotechnology, Inc. (B.C., P.P., P.G.-D., J.M., J.S., D.H., B.G., F.F.), by the Hartford and McKnight Foundations (I.L.) and the NIH (I.L., K.D., W.W.). The sequence data in this publication have been submitted to the EMBL/GenBank Libraries under the accession number Y00297.

Received 25 September; accepted 16 December 1987.

1. Masters, C. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 4245–4249 (1985).
2. Glenner, G. G. & Wong, C. W. *Biochem. biophys. Res. Commun.* **120**, 885–890 (1984).
3. Glenner, G. G. & Wong, C. W. *Biochem. biophys. Res. Commun.* **122**, 1131–1135 (1984).
4. Kang, J. *et al. Nature* **325**, 733–736 (1987).
5. Goldgaber, D., Lerman, M. I., McBride, O. W., Saffiotti, U. & Gajdusek, D. C. *Science* **235**, 877–880 (1987).
6. Tanzi, R. E. *et al. Science* **235**, 880–884 (1987).
7. Robakis, N. K., Ramakrishna, N., Wolfe, G. & Wisniewski, H. M. *Proc. natn. Acad. Sci. U.S.A.* **84**, 4190–4194 (1987).
8. St. George-Hyslop, P. H. *et al. Science* **238**, 664–666 (1987).
9. Tanzi, R. E., Bird, E. D., Latt, S. A. & Neve, R. L. *Science* **238**, 666–669 (1987).
10. Podlisny, M. B., Lee, G. & Selkoe, D. J. *Science* **238**, 669–671 (1987).
11. Nawa, H., Kotani, H. & Nakanishi, S. *Nature* **312**, 729–734 (1984).
12. Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S. & Evans, R. M. *Nature* **298**, 240–244 (1982).
13. Martinez, R., Mathey-Prevot, B., Bernards, A. & Baltimore, D. *Science* **237**, 411–416 (1987).
14. Anderson, S. & Kingston, I. B. *Proc. natn. Acad. Sci. U.S.A.* **80**, 6838–6842 (1983).
15. Gebhard, W. & Hochstrasser, K. in *Proteinase Inhibitors* (eds Barrett, A. J. & Salvesen I. G.) 389–402 (Elsevier, Amsterdam, 1986).
16. Travis, J. & Salvesen, G. *A. Rev. Biochem.* **52**, 655–709 (1982).
17. Wolf, D. & Rotter, V. *Proc. natn. Acad. Sci. U.S.A.* **82**, 790–794 (1985).
18. Benton, W. D. & Davis, R. W. *Science* **196**, 180–182 (1977).
19. Maniatis, T., Fritsch, E. F. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, New York, 1982).
20. Rigby, P. W., Dieckmann, M., Rhodes, C. & Berg, P. *J. molec. Biol.* **113**, 237–251 (1977).
21. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).
22. Lipman, D. J. & Pearson, W. R. *Science* **227**, 1435–1441 (1985).
23. Laskowski, M. & Kato, I. *A. Rev. Biochem.* **49**, 593–626 (1980).
24. Cathala, G. *et al. DNA* **2**, 329–335 (1983).
25. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. *Biochemistry* **18**, 5294–5299 (1979).
26. Aviv, H. & Leder, P. *Proc. natn. Acad. Sci. U.S.A.* **69**, 1408–1412 (1972).
27. Thomas, P. *Proc. natn. Acad. Sci. U.S.A.* **77**, 5201–5205 (1980).
28. Deng, G. & Wu, R. *Nucleic Acids Res.* **9**, 4173–4176 (1981).
29. Gunning, P. *et al. Molec. Cell Biol.* **3**, 787–795 (1983).
30. Kassell, R. & Laskowski, M. *Biochem. biophys. Res. Commun.* **20**, 463–468 (1965).
31. Wachter, B., Deppner, K., Hochstrasser, K., Lempart, K. & Geiger, R. *FEBS Lett.* **119**, 58–62 (1980).
32. Cechova, D., Jonakova, V. & Sorm, F. *Collect Czech. Chem. Commun.* **36**, 3342–3357 (1971).
33. Wachter, E. & Hochstrasser, K. *Hoppe-Seyler's Z. physiol. Chem.* **362**, 1351–1355 (1981).
34. Hochstrasser, K. & Wachter, E. *Hoppe-Seyler's Z. physiol. Chem.* **364**, 1679–1687 (1983).

Exhibit D

1802–1808 *Nucleic Acids Research, 1997, Vol. 25, No. 9*

A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus

Masahira Hattori[1,*], Fujiko Tsukahara[1,2], Yoshiaki Furuhata[1], Hiroshi Tanahashi[1], Matsumi Hirose[1], Masae Saito[1], Shiho Tsukuni[1] and Yoshiyuki Sakaki[1]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108, Japan and [2]Department of Pharmacology, Tokyo Woman's Medical College, Tokyo 162, Japan

Received December 16, 1996; Revised and Accepted March 24, 1997

DDBJ/EMBL/GenBank accession no. D87675

ABSTRACT

We developed a novel *in vitro* method for making nested deletions and applied it to a large-scale DNA sequencing. A DNA fragment to be sequenced (up to 15 kb long) was cloned with a new vector possessing two unique *Sfi*I sites, digested by *Sfi*I and ligated to generate a large head-to-tail concatemer. The large concatemer was randomly fragmented by sonication and then redigested by *Sfi*I to separate insert and vector DNAs. The fragments of various length were then cloned into the other vector(s) specifically designed for selective cloning of insert-derived DNA fragments to generate a library of nested deletions. This method allowed a single person to generate >20 nested deletion libraries sufficient to cover 100 kb in a few days. We applied the method for sequencing of P1 clones and successfully determined the complete sequence of ~300 kb of the human amyloid precursor protein (APP) locus on chromosome 21 with a redundancy of 3.8, reasonably low cost and very few gaps remaining to be closed. Development of some new instruments and software is also described which makes this method more applicable for large-scale sequencing.

INTRODUCTION

Large-scale sequencing is now one of the central issues in the human genome project (1,2). For sequencing the genome with reasonable speed, accuracy and cost, sequencing strategy is an important factor to be considered. So far, three types of strategies have been used or proposed, namely, shot-gun, primer-walk and nested deletion strategies (reviewed in ref. 3). Each strategy has advantages and disadvantages in practical use. The shot-gun strategy has been most widely used for large-scale sequencing projects (i.e., 4,5). It is simple as a whole and easy to scale up. However, intrinsically it requires a high redundancy of sequencing and extensive gap closure efforts, which may create cost and data assembly process problems, respectively. Primer-walk is a directed strategy with a minimum redundancy but it requires the design and synthesis of a very large number of primers, which may be expensive. It has been proposed to use a library of short oligonucleotides as sequencing primers (6,7) but it has not been technically well-established. The third strategy is sequential sequencing by using nested deletion or transposon-inserted templates (8). This strategy could be carried out with reasonably low redundancy and simple data assembly process, but has not been considered to be applicable for large-scale sequencing because of its complicated procedure for template preparation. However, a relatively simple, transposon-mediated method has been developed and successfully applied for sequencing of the *Drosophila* genome (9).

We have attempted to develop a simple and reproducible method for making nested deletions on a large-scale. We herein describe the novel method for making nested deletions *in vitro* and its successful application for sequencing ~300 kb of human APP locus on chromosome 21q22.1. Development of some instruments and software is also described, which makes this method applicable for the large-scale and systematic sequencing of the human genome.

MATERIALS AND METHODS

Construction of pSFI vectors

Oligonucleotides for polylinker were prepared by a DNA synthesizer (Perkin-Elmer ABI 394). The double-stranded DNAs described below (a, b and c) were prepared by annealing an equimolar of the synthesized complementary oligonucleotides at 50°C overnight in 100 µl of 0.1 M NaCl. The three double-stranded DNAs having 5′ overhang cohesive ends of *Hin*dIII and *Eco*RI sites were ligated with *Hin*dIII and *Eco*RI double-digested pUC13 to construct pSFI-CV (a) or pTZ19R for pSFI-SV1 (b) and pSFI-SV10 (c). The polylinker of pSFI-CV contains *Eag*I, *Sal*I, *Hin*dIII, *Bam*HI and two *Sfi*I sites. The polylinker of pSFI-SV1 contains *Hin*dIII, *Sfi*I, *Bgl*II, *Eco*RV and *Eco*RI sites and the polylinker of pSFI-SV1 contains *Hin*dIII, *Sfi*I, *Bgl*II, *Stu*I and *Eco*RI sites, respectively.

[*]To whom correspondence should be addressed. Tel: +81 3 5449 5623; Fax: +81 3 5449 5445; Email: hattori@hgc.ims.u-tokyo.ac.jp

```
(a) 5′-AGCTGGCCAAATCGGCCGTCGACAAGCTTGGATCCGGCCATAAGGGCC
          CCGGTTTAGCCGGCAGCTGTTCGAACCTAGGCCGGTATTCCCGGTTAA-5′
(b) 5′-AGCTTGCATGCCAGGCCAAATCGGCCCTAGGAGATCTGATATCAGGCCTGAGCTCG
          ACGTACGGTCCGGTTTAGCCGGCTTCCTCTAGACTATAGTGGGACTCGAGCTTAA-5′
(c) 5′-AGCTTGCATGCCAGGCCATTAGGGCCGAGATCTGGAGGCCTCCCGGGGAGCTCG
          ACGTACGGTCCGGTAATCCCGGCTCTAGACCTCCGGAGGGCCCCTCGAGCTTAA-5′
```

Kanamycin-resistant pSFI-SV vectors were further prepared by ligation of the *Bsp*HI-digested plasmid having the new polylinker with the *Bam*HI fragment containing kanamycin-resistant gene of pBS-Kan2 (10). For preparation of nested deletions, the pSFI-SV1 and -SV10 were double-digested by *Eco*RV and *Sfi*I or *Stu*I and *Sfi*I, treated by CIP and gel-purified. The pSFI-CV was digested with appropriate restriction enzymes which cleave the multiple cloning site, treated by CIP and gel-purified. The MCSs of these three vectors are in-frame for the *Escherichia coli Lac* Z gene which produces blue colonies in the presence of X-*gal*. The structure of these vectors is shown in Figure 1.

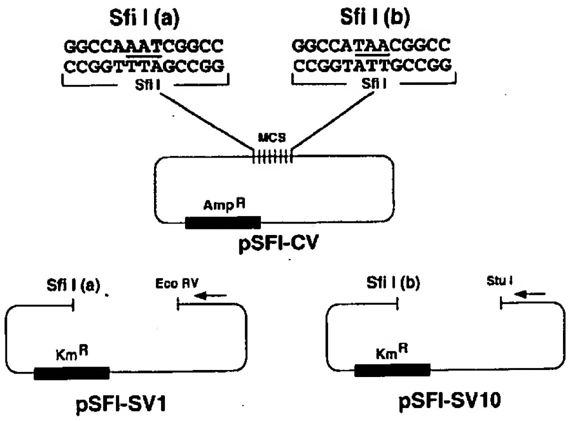

Figure 1. Vectors for construction of nested deletions. The pSFI-CV contains a multiple cloning site (MCS) for *Eag*I, *Sal*I, *Hin*dIII and *Bam*HI, which is used for cloning of foreign DNA fragments to be sequenced. The MCS is flanked by two *Sfi*I sites having the overhang ends of AAT-3′ (a) and TTA-3′ (b), respectively. Two pSFI-SV vectors (SV1 and SV10) were constructed for cloning of nested deletions generated from pSFI-CV clone. The pSFI-SV1 contains a *Sfi*I site producing the overhang end of AAT-3′ and a blunt-end of *Eco*RV. The pSFI-SV10 contains a *Sfi*I site producing the overhang end of TTA-3′ and a blunt-end of *Stu*I. Each *Sfi*I end of the pSFI-SV vectors is complementary to either of the *Sfi*I ends of the insert DNA in the pSFI-CV. Amp^R and Km^R indicate ampicillin and kanamycin resistant genes, respectively. Arrows indicate the direction of sequencing.

Preparation of nested deletion library

The overall procedure for the preparation of nested deletions is illustrated in Figure 2. The pSFI-CV clone was cultured overnight in the presence of ampicillin and the plasmid was isolated by the alkaline-SDS method (11). The plasmid DNA (20 µg) was digested with *Sfi*I (NEB, 40 U) in a final volume of 100 µl at 50°C for 1 h and extracted by phenol/chloroform and precipitated by ethanol. The DNA was treated with T4 DNA ligase (Takara, 10 U) and 0.1 mM ATP under DNA concentrations of 0.5–1 µg/µl at 15°C in 40 µl of 1× ligation buffer for 2 h to overnight. An aliquot of the viscous ligated mixture was diluted to ~5 ng/µl in 200 µl TE (10 mM Tris–HCl, 1 mM EDTA, pH 8.0) and sonicated by a sonicator Astrason XL with the pulsar dial 1.5–2.0. The time for sonication depends on the insert DNA size and was routinely set for 30 s–1 min for 5 kb DNA, in which the broad range of smear bands were obtained. The sonicated DNA was extracted by phenol/chloroform, precipitated by ethanol and dissolved in 30 µl TE. The DNA was treated with T4 DNA polymerase (Toyobo, 0.5 U) at 37°C for 5 min in a 20 µl of 50 mM Tris–HCl, pH 7.5, 10 mM $MgCl_2$, 10 mM DTT and 0.2 mM dNTPs. The reaction was quenched by heating at 70°C for 15 min, then the DNA was treated with T4 polynucleotide kinase (Takara, 5 U) in the presence of 0.1 mM ATP in 30 µl of appropriate buffer at 37°C for 20 min. The DNA was further digested with *Sfi*I (NEB, 10 U) in 80 µl of appropriate buffer at 50°C for 1 h. The reaction was quenched by adding EDTA and the DNA was extracted by phenol/chloroform, precipitated with ethanol and dissolved in 20 µl TE. An aliquot of the digested DNA (0.1–0.3 µg) was ligated with pSFI-SV1 or pSFI-SV10 (0.1 µg) in 20 µl of 1× ligation buffer containing 4 U of T4 DNA ligase (Takara) and 0.1 mM ATP at 15°C for 2 h to overnight. The ligated mixture was transformed to *E.coli* DH5α (Gibco-BRL) and the colonies resistant to kanamycin were obtained as nested deletion library.

Size measurement and ordering of nested deletions

The kanamycin resistant colonies were randomly picked and cultured in 5 ml L-broth in the presence of kanamycin (100 µg/ml) overnight and the plasmid DNA were isolated by an automated plasmid isolator PI100 (KURABO, Japan). Or the insert DNA was directly amplified from colonies by PCR (long PCR kit from Takara or Gibco-BRL) using the specific primers (LR: 5′-TCCGGCTCGTATGTTGTGTGGA-3′, LL: 5′-GTGCTGCAAGGCGATTAAGTTGG-3′) for 30 cycles of 94°C for 30 s and 68°C for 1–15 min followed by one cycle at 70°C for 10 min. The plasmid DNAs digested with *Sfi*I or the PCR-amplified products were electrophoresed on a 0.8% agarose gel and the gel image was recorded by using a CCD imaging system (ATTO Co. Ltd., Tokyo, Japan). The size measurement and ordering of the clones were performed by a computer software program Lane Screener newly developed for this purpose by ATTO. The nested deletion clones were selected at the interval of 250–350 bases for sequencing. The nested deletion plasmids were used for sequencing without further purification or the PCR products were used after treating with shrimp alkaline phosphatase–*E.coli* exonuclease I (Amersham) at 37°C for 20 min followed by heating at 85°C for 10 min.

Sequencing and data assembly of nested deletions

Sequencing was done by cycle sequencing with a commercially available fluorescent-labeled forward primer (–21m) and analyzed by a four color-based sequencer (Perkin-Elmer ABI 373S). The sequencing reaction was carried out by a manual manner in the beginning and later by a sequencing robot (Vistra, Amersham) according to the manufacturer's instructions. The buffers and reagents for sequencing were obtained from the manufacturers. Data assembly was done by a commercially available program ATSQ (Japan Software Inc., Tokyo, Japan) in the beginning and later by a newly developed system SAND (see Results).

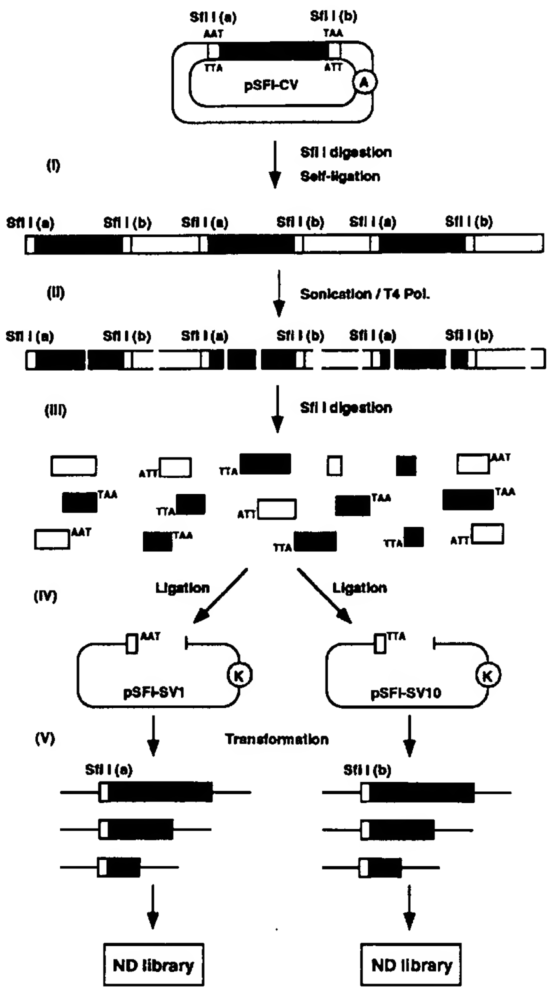

Figure 2. Procedure for construction of nested deletions. The nested deletions are constructed by five steps. (I) The pSFI-CV clone is digested by *Sfi*I and then the fragments are ligated to obtain alternatively multimerized high molecular weight DNA. (II) The ligated DNA is sonicated to obtain the fragments of various size and then flushed by treatment of T4 DNA polymerase. (III) The sonicated DNA is digested by *Sfi*I to obtain the fragments with a *Sfi*I end at the one end and a blunt-end at the other end. (IV) Finally, the *Sfi*I-digested fragments are ligated to either of pSFI-SV vectors and (V) introduced into *E.coli* to obtain the nested deletion library. Open and closed boxes indicate vector and cloned DNA fragments, respectively. Capital A and K in circles indicate ampicillin and kanamycin resistant genes, respectively.

Subcloning of restriction fragments from P1 DNA

The P1 clone was cultured in 100 ml of L-broth in the presence of kanamycin (100 μg/ml) and the DNA was isolated by alkaline-SDS method according to the literature (14). The crude P1 DNA was treated with RNase A (Sigma, 20 μg) at 37°C for 1 h and precipitated by adding 0.6 vol of 2.5 M NaCl/20% polyethylene-glycol 6000. The mixture was kept on ice for 15 min and the DNA was precipitated by centrifugation. The precipitate was rinsed by 75% ethanol, dried and dissolved in 100 μl TE. An aliquot (20 μl) was digested by *Avr*II, *Xba*I, *Bam*HI or *Bgl*II for 1 h and then treated with Klenow fragment (0.1 U) in the presence of 0.2 mM dCTP/TTP or dGTP/dATP to partially fill-in the restriction end at 37°C for a further 15 min. The reaction was quenched by adding 1 μl of 0.5 M EDTA and the DNA was extracted with phenol/chloroform, precipitated by ethanol and dissolved in 20 μl TE (pH 8). An aliquot (1–5 μl) was ligated with a pSFI-CV (0.1 μg) partially filled-in at the *Hin*dIII, or *Sal*I site in 20 μl of 1× ligation buffer containing 4 U T4 DNA ligase (Takara) and 0.1 mM ATP at 15°C for 2 h to overnight. The ligation mixture was transformed into *E.coli* DH5α (Gibco-BRL) and the colonies resistant to ampicillin were obtained as subclones. Sixty colonies were randomly picked and cultured in 5 ml of L-broth in the presence of ampicillin (100 μg/ml) overnight and the plasmid DNA was isolated by alkaline-SDS method and dissolved in 150 μl of TE. These clones were stored as the sub-libraries and subjected to the following fingerprinting selection. The cloned DNA (10 μl) was digested with *Hae*III (Nippon gene, 4 U) at 37°C for 1 h and electrophoresed on a 2% Nusieve/1% agarose gel. The *Hae*III fingerprint of each clone was used to choose the independent subclones. The size of the insert DNA was also estimated by digestion with *Sfi*I. The sequences of the both-ends of each subclone were determined by cycle sequencing with forward (endash 21m) and reverse (M13 RP) primers.

RESULTS

Construction of the nested deletion libraries

To construct nested deletion libraries, we designed and constructed three unique vectors, termed pSFI-CV and pSFI-SV1 and pSFI-SV10, from commercially available plasmid vectors pUC13 and pTZ19R as described in Materials and Methods. As shown in Figure 1, pSFI-CV has a multiple cloning site (MCS) which is sandwiched by the two unique *Sfi*I sites producing the 3′-overhang ends of ATT-3′ and TAA-3′, respectively. Two pSFI-SV vectors have a MCS producing a blunt-end and a *Sfi*I 3′-overhang end complementary to either *Sfi*I end of the pSFI-CV, respectively. These vectors are available upon request. By using these vectors, nested deletion libraries were generated as follows (Fig. 2). A DNA fragment to be sequenced (usually 2–15 kb long) was cloned into the MCS of pSFI-CV. The plasmid was digested with *Sfi*I and then ligated at a high DNA concentration to generate the alternatively ligated high-molecular weight DNA. The ligated DNA was sonicated to generate the various sizes of fragments and the ends were flushed by the treatment with T4 DNA polymerase. The sonication of high molecular weight DNA enabled us to generate various size of fragments more efficiently than that of un-treated small plasmid DNA. The blunt-ended DNA was digested with *Sfi*I to produce the fragments with a blunt-end at one end and a *Sfi*I 3′-overhang end at the other end. Among the fragments, only the insert-derived fragments have the *Sfi*I site complementary to the *Sfi*I end of either pSFI-SV vector. The fragments were ligated with a pSFI-SV1 or SV10 double-digested with *Sfi*I and appropriate blunt-end enzymes, which enabled us to selectively ligate the insert DNA-derived fragments with pSFI-SV1 or pSFI-SV10 through the complementary end of the *Sfi*I. Finally, the ligated DNAs were introduced into *E.coli* to generate a library of the nested deletion clones of various length. The results of construction of nested deletions from several fragments are summarized in Table 1. The data indicated that ~90% of the clones in the library were nested deletions with the various sizes. In general, smaller DNAs were preferentially cloned but some uneven size distribution of the deletions caused little problem in practical use.

Table 1. Construction of nested deletions from various size of DNA fragments

Clone	Estimated insert size (kb)	Vector	No. of isolated clones	No. of nested deletions	Yield of nested deletions (%)	Size range of insert DNA (kb)
1	2.5	SV1	40	36	90	0.2–2.3
		SV10	40	35	87	0.3–2.3
2	3.6	SV1	60	52	87	0.1–3.7
		SV10	60	43	71	0.2–3.3
3	4.0	SV1	60	58	97	0.3–3.9
		SV10	60	54	90	0.3–3.9
4	6.0	SV1	90	58	64	0.2–5.9
		SV10	90	71	80	0.4–6.0
5	7.2	SV1	96	82	85	0.2–6.7
		SV10	96	65	68	0.4–6.8
6	8.5	SV1	120	107	89	0.2–7.6
		SV10	120	100	83	0.1–8.0
7	11.3	SV1	120	113	86	0.1–10.4
		SV10	120	119	99	0.4–10.8

Insert DNA was amplified by colony PCR except clones 6 and 7, which were isolated as plasmids. The DNA was subjected to electrophoresis on a 0.8% agarose gel and the size was measured by CCD imaging system. Clones other than nested deletions were un-amplified or had no insert.

Isolation and ordering of nested deletion clones

The kanamycin-resistant white colonies prepared as above were randomly picked, cultured and the plasmids were isolated by an automated plasmid isolator. Alternatively, the insert DNA was directly amplified from colonies by PCR. The plasmid or PCR-amplified DNA was electrophoresed on agarose gel for size measurement. Figure 3a shows a typical example of the electrophoresis pattern of PCR-amplified insert DNAs of randomly isolated clones from a nested deletion library. Since the size measurement is a laborious step, we developed an automated size-measuring system, in collaboration with ATTO Co. Ltd. (Tokyo, Japan). The system allowed us to automatically measure the size of insert DNAs and to re-align the clones by size with the aid of the computer program, Lane Screener. The CCD Imaging system and the program Lane Screener will be purchased from ATTO. Figure 3b shows an example of the electrophoresis pattern of the ordered nested deletions. Among ordered clones, appropriate clones were chosen with appropriate deletion interval and subjected to sequencing. Since 400–600 bases of raw data can be constantly obtained by sequencing of one clone, a set of nested deletions with an interval of 250–350 bases is sufficient to obtain the contiguous sequence data of one strand. For example, in the case of a 4 kb DNA fragment, 12 clones out of the 60 ordered clones for one strand were selected for sequencing.

Sequencing and data assembly

The selected clones were subjected to sequencing analysis. The sequencing was carried out as described in Materials and Methods. The data assembly was initially performed by a commercially available software program ATSQ (Japan Sofware Inc., Japan) developed for shot-gun sequencing. However, our strategy gives an ordered sequencing data that can be assembled without the unnecessary matching process of positionally unrelated sequences. We thus developed a new data assembly program for the nested deletion strategy in collaboration with Mitsui Knowledge Industry Co., Ltd. (Japan). The system named Sequence Assembler of Nested Deletion (SAND) enabled us to assemble the data with high speed and accuracy. Details of SAND will be published elsewhere.

Application of the nested deletion strategy for large-scale sequencing

We applied our nested deletion method for sequencing ~300 kb of the human amyloid precursor protein (APP) locus mapped on chromosome band 21q22.1. All the exons and their flanking regions of the APP gene were previously isolated and partly sequenced in our laboratory (12). For this study, we isolated five P1 clones covering the entire APP gene (Fig. 4) from a P1 library specific for human chromosome 21q (13).

The overall procedure for sequencing of the P1 clones consisted of: (i) subcloning of restriction fragments of the P1 clone into the pSFI-CV; (ii) construction of the restriction map of the P1 clone; (iii) construction of nested deletions from selected subclones; (iv) sequencing and data assembly of the nested deletions; and (v) assembly of the consensus sequence data of each subclone to generate the final contiguous sequence data.

We first subcloned the restriction fragments from each P1 clone into pSFI-CV. We used *Bam*HI, *Xba*I, *Avr*II and *Bgl*II, because *Bam*HI, *Xba*I, *Avr*II do not appear and *Bgl*II appears once in the P1 vector (pAd10SacIIB, ref. 14). The restriction subclones were randomly isolated from a P1 clone and briefly characterized by *Hae*III-fingerprinting to choose independent clones. Among the 240 subclones, ~80 of independent clones with a range of 0.3–16 kb in size were chosen. The both-end sequencing of the independent clones followed by the data assembly were performed to obtain the partial map. This process usually produced three to five contiguous regions of 10–40 kb. The gaps between the contigs

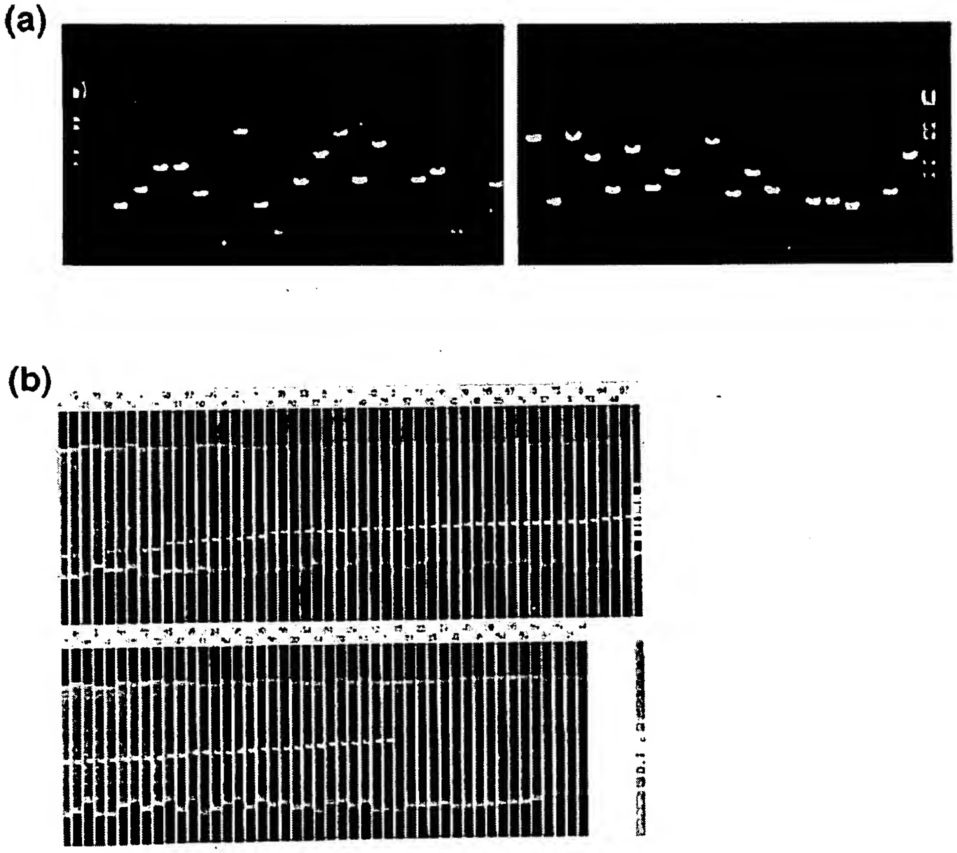

Figure 3. An example of size-measurement and ordering of nested deletions. (a) Electrophoresis profiles of the PCR-amplified DNAs. The insert DNA was originally 4.6 kb long and the amplification product of no insert clone is ~200 bp. The most left lane of the left gel and the most right lane of the right gel indicate a size marker, *Sty*I-digested λ DNA. The size of each band is ~19.3, 7.7, 6.2, 4.3, 3.5, 2.7, 1.9, 1.5, 0.9 and 0.4 kb from the top, respectively. (b) The size of DNA band was measured by the CCD imaging system (ATTO Co. Ltd., Japan), and the bands were ordered by the size by newly developed software program, Lane Screener, which enables us to simultaneously measure and order up to 300 fragments and also to select the clones with appropriate deletion-interval for sequencing. The process from size-measurement to selection of nested deletions for sequencing can be automatically done in 15 min for 100 fragments.

were then filled by re-screening of remaining subclones by PCR using primers designed from the sequence data of the edges of each contig. This procedure allowed us to rapidly construct the restriction map and to construct the minimum tiling path of the fragments over the five contiguous P1 clones. Finally, we selected 72 restriction fragments from the five P1s. The nested deletion libraries were prepared for 69 restriction fragments and finally ~1600 of nested deletions were subjected to sequencing analysis. The remaining three fragments of <1.5 kb were sequenced by primer-walk method using dye-terminator chemistry. We finally obtained 301 692 bp of the consensus sequence (DDBJ accession number; D87675) from the five P1 clones. The results revealed the complete structure of the APP gene spanning 286 722 bp (Fig. 4). Detailed analysis of the sequence data will be reported elsewhere. Table 2 summarized the efficiency of our method for sequencing of the APP locus. It should be noted that the final data were obtained with a redundancy of 3.8 that is about a half of that by shot-gun strategy. The typical shot gun strategy would require redundancy in a range from 6 to 8 (15). The redundancy by primer-walk strategy would be estimated at ~2.5, assuming that a reaction produces a 500 base data and the 100 bases are overlapped between the data. But this value is an underestimation for sequencing of human genome because the presence of highly repetitive sequences such as the *Alu* family makes it difficult to continue the walk. It should be also noted that gap closure for only six gaps was required for the 300 kb sequencing. These gaps were due to the failure of sequencing of the PCR-amplified templates containing poly-purine, poly-pyrimidine or poly purine-pyrimidine sequences and were successfully closed by re-sequencing the original subclones as templates by dye-terminator chemistry using specific primers. About 96% of double-stranded coverage was obtained at the initial step and the remaining 4% (29 regions) of single-stranded coverage were caused by missing of nested deletions from either strand. These regions were also covered by sequencing the original subclones by primer-walk method using dye-terminator chemistry. In total, 44 synthetic primers were required to complete the double-stranded coverage of 300 kb.

DISCUSSION

In the present study, we developed a novel method for making nested deletions and demonstrated that it is applicable for

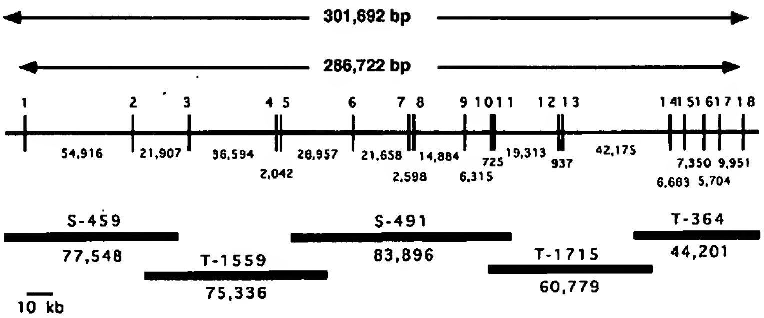

Figure 4. P1 clones covering the human APP locus and exon–intron organization of the APP gene based on the sequence data obtained in this study. The five overlapping P1 clones (S-459, T-1559, S-491, T-1715 and T-364) were used as starting materials for sequencing. The position of the exons (1–18) was arbitrarily shown by vertical bars. The numbers between the exons indicate the size of the introns in bp. The sequence data revealed that the five P1 clones covered 301 692 bp. The first exon started at 9001 and the end of the last exon (exon 18) located at 295 722, so that the APP gene is 286 722 bp in length. The accession number of the entire sequence of the APP locus is D87675.

large-scale sequencing. The method for making nested deletions is comprised of simple steps including digestion by *Sfi*I, ligation and sonication. It has not yet been automated but is simple enough to obtain the nested deletion libraries for >100 kb of final sequence data in a few days by one technician. The use of asymmetric 3′-overhang ends of *Sfi*I cleavage sites in the vector provides a high efficiency and selectivity in the ligation steps. Actually, the nested deletion libraries prepared by this method contained almost no vector-derived fragments (see Table 1). The method is applicable for all the clonable DNA, although some modifications may be required for the DNA fragments having *Sfi*I sites of the same 3′-overhang ends with those of pSFI vectors. On average, *Sfi*I site appears every 200 kb in the human genome and the *Sfi*I sites with the same 3′-overhang end every 13 000 kb or less, thus the frequency of such *Sfi*I sites is low enough for practical use.

To apply this method for sequencing of the large insert clones such as P1 and BAC clones, subcloning and mapping of the restriction fragments are required. In the case of P1 clones, this step could be simply performed by one-pass end-sequencing of the restriction fragments followed by the data assembly on the computer, and provided several advantages for practical use. It made the final data assembly easy. It was also useful to find the same fragments present in the overlapped regions between the P1 clones to avoid unnecessary duplication of sequencing. Also, the use of restriction enzyme allowed us to use P1 DNA purified simply by polyethylene-glycol precipitation in the subcloning step. No ultra-centrifugation step was required.

The most labor-intensive and time-consuming step of the nested deletion-based strategy is the isolation and ordering of nested deletions. For example, we isolated and electrophoresed ~8000 clones for selecting 1600 nested deletions. However, the employment of the direct amplification of insert DNA by long PCR (16) enabled us to rapidly prepare templates of various lengths. The templates prepared in this manner gave sequence data of high quality sufficient for the data assembly. For ordering the clones, we developed a new instrument for automated size-measurement and ordering of the clones, which allowed us to handle >2000 nested deletion clones in a day per person. A software program was also developed for automated selection of nested deletions with appropriate deletion interval.

Table 2. Summary of sequencing of APP locus using the nested deletion method

Description	Number
P1 clones sequenced	5 clones
Independent restriction fragments	368 fragments
Restriction fragments sequenced	72 fragments
Restriction fragments used for ND construction	69 fragments
Restriction fragments used for primer-walk	3 fragments
ND clones isolated	8120 clones
Total sequencing reactions	2370 reactions
End-sequencing of restriction fragments	736 reactions
Sequencing of small fragments by primer-walk	8 reactions
Sequencing of ND clones	1582 reactions
Primer-walk for gap closure	12 reactions
Primer-walk for double-stranded coverage	32 reactions
Average edited read base	486 bases
Total read base	1 151 820 bases
Consensus sequence	301 692 bases
Redundancy	3.8

ND, nested deletions.

In contrast to relatively complicated steps in the front side, the steps for sequencing and data assembly are easy and simple. To obtain the final 300 kb contiguous data of the APP locus, only 2400 sequencing reactions including end-sequencing of the subclones and gap-close sequencing were carried out in total. The data were quickly assembled in a quite simple and reliable manner. The results were easily checked by comparing the clone alignment in the assembled data with those obtained from size-measurement. It should be emphasized that in our method the direction of each sequence data was already known prior to data assembly, so that at first the data for each strand were assembled and then the data of each strand were compared to obtain the final consensus data. This two-step assembly procedure eliminated the unnecessary matching process of the complementary sequence data of each sample, accelerating the overall speed of the data assembly step and making it easier to resolve the ambiguities.

The accuracy of the data obtained in this study is estimated to be >99.99% as discussed below. We compared raw data from 1000 samples and found that the frequency of ambiguity of our data was 0.25% in bases 1–400 and ~10% in bases 401–500. The data assembly for one strand was carried out by using up to 500 bases of raw data and the ambiguities were resolved by eye inspection. Since we used the data with an interval of ~300 bases, 400 bases out of the 500 bases of raw data are overlapped each other and the ambiguity in the overlapped region is estimated to 0.005% (0.25% × 2% = 0.005%). The accuracy of the remaining 100 bases of non-overlapped region is 0.25% as mentioned above. Therefore, the assembled data for one strand should have an accuracy of at least 99.75%. However, for practical purposes, the ambiguities were finally resolved based on the data from both-stranded coverage over the entire region. Therefore, the error rate should be much less than 0.01% (0.25% × 0.25% = 0.000625%). Also, the final consensus data were obtained with a redundancy of 3.8. Thus, the accuracy should be much higher than the above value. In fact, comparison of the present data with those of 3.6 kb APP cDNA showed no inconsistency between them.

In conclusion, the nested deletion method developed in this study provided a novel strategy for large-scale sequencing characterized by low redundancy, little gap closure effort and rapid data assembly. The procedure can be employed not only for P1 but also for BAC (17) and PAC (18) systems.

ACKNOWLEDGEMENTS

This work was supported in part by Grant-in-Aid for a Creative Basic Research (Human Genome Program) from the Ministry of Education, Science, Sports and Culture, and a Grant of a Special Coordination Funds for Promotion of Science and Technology from the Science and Technology Agency (STA) of Japan.

REFERENCES

1 Gibbs, R. A. (1995) *Nature Genet.*, 11, 121–125.
2 Olson, M. V. (1995) *Science*, 270, 394–396.
3 Hunkapiller, T., Kaiser, R. J., Koop, B. F. and Hood, L. (1991) *Science*, 254, 59–67.
4 Hodgkin, J., Plasterk, R. H. and Waterston, R. H. (1995) *Science*, 270, 410–414 .
5 Fleischmann, R. D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269, 496.
6 Studier, F. W. A (1989) *Proc. Natl. Acad. Sci. USA*, 86, 6917–6921.
7 Kotler, L. E., Zevin-Sonkin, D., Sobolev, I. A., Beskin, A. D. and Ulanovsky, L. E. (1993) *Proc. Natl. Acad. Sci. USA*, 90, 4241–4245.
8 Strathmann, M., Hamilton, B. A., Mayeda, C. A., Simon, M. I., Meyerowitz, E. M. and Palazzolo, M. J. (1991) *Proc. Natl. Acad. Sci. USA*, 88, 1247–1250.
9 Martin, C. H., Mayeda, C. A., Davis,C. A., Ericsson, C. L., Knafels, J. D., Mathog, D. R., Celniker, S. E., Lewis, E. B. and Palazzolo, M. J. (1995) *Proc. Natl. Acad. Sci. USA*, 92, 8398–8402.
10 Kusuda, J. Kameoka, Y. Takahashi, I. Fujiwara, H. and Hashimoto, K. (1989) *Nucleic Acids Res.*, 17, 8890.
11 Maniatis, T., Fritsch, E. F. and Sambrook, J. (1989) *Molecular cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
12 Yoshikai, S., Sasaki, H., Doh-ura, K., Furuya, H. and Sakaki, Y. (1990) *Gene*, 87, 257–263.
13 Tanahashi, H., Ito, T., Hattori, M., Ohira, M., Ohki, M., Tashiro, K. and Sakaki, Y. (1994) *DNA Res.*, 1, 85–89.
14 Pierce, J. C. Sauer, B. and Sternberg, N. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 2056–2060.
15 Hodgkin, J., Plasterk, R. H. A. and Waterston, R. H. (1995) *Science*, 270, 410–414.
16 Barnes, W. M. (1994) *Proc. Natl. Acad. Sci. USA*, 91, 2216–2220.
17 Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992) *Proc Natl. Acad. Sci. USA*, 89, 8794–8797.
18 Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A. and De Jong, P. J. (1994) *Nature Genet.*, 6, 84–89.

Exhibit E

NCBI Entrez Protein My N[Sign In] [Regis

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features: ☑ CDD Refresh

□ 1: CAA31830. Reports A4 amyloid protei...[gi:871360]

BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       871360                   695 aa            linear       20-JUN-1995
DEFINITION  A4 amyloid protein precursor.
ACCESSION
VERSION       GI:871360
DBSOURCE    embl locus HSPRA41, accession X13466.6 release 44
            accession X13467
            accession X13468
            accession X13469
            accession X13470
            accession X13471
            accession X13472
            accession X13473
            accession X13474
            accession X13475
            accession X13476
            accession X13477
            accession X13478
            accession X13479
            accession X13487
            accession X13488
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
REFERENCE   1  (residues 1 to 695)
  AUTHORS   Lemaire,H.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-NOV-1988) Lemaire H.G., Institut fuer Genetik, der
            Universiteit zu Koeln, Weyertal 121, D-5000 Koeln 41, FRG
REFERENCE   2  (residues 1 to 695)
  AUTHORS   Lemaire,H.G., Salbaum,J.M., Multhaup,G., Kang,J., Bayney,R.M.,
            Unterbeck,A., Beyreuther,K. and Muller-Hill,B.
  TITLE     The PreA4(695) precursor protein of Alzheimer's disease A4 amyloid
            is encoded by 16 exons
  JOURNAL   Nucleic Acids Res. 17 (2), 517-522 (1989)
  MEDLINE   89128427
COMMENT     On Jun 23, 1995 this sequence version replaced gi:35600.
FEATURES             Location/Qualifiers
     source          1..695
                     /organism="unknown"
     Protein         1..695
                     /product="A4 amyloid protein precursor"
     Region          24..188
                     /region_name="A4_EXTRA"
                     /note="amyloid A4; amyloid A4 precursor of Alzheimers
                     disease; smart00006"
                     /db_xref="CDD:47362"
     Region          <309..505
                     /region_name="SbcC"
                     /note="ATPase involved in DNA repair [DNA replication,
                     recombination, and repair]; COG0419"
                     /db_xref="CDD:30768"
     Region          600..>630
                     /region_name="Beta-APP"
                     /note="Beta-amyloid peptide (beta-APP); pfam03494"
                     /db_xref="CDD:43420"
     CDS             1..695
                     /gene="PreA4"
                     /coded_by="join(X13466:1..57,X13467.1:452..619,
                     X13468.1:12..141,X13469.1:89..201,X13470.1:116..309,
                     X13471.1:1..203,X13472.1:85..218,X13473.1:52..126,
                     X13474.1:160..318,X13475.1:1..129,X13476.1:9..108,
                     X13477.1:193..414,X13478.1:1..54,X13479.1:137..237,
                     X13487.1:179..325,X13488.1:171..272)"
                     /note="pid:e"
                     /label=preA4_CDS
ORIGIN
        1 mlpglallll aawtvwalev ptdgnaglla epqiamfcgr lnmhmnvqng kwdsdpsgtk
       61 tcidtkegil qycqevypel qitnvveanq pvtiqnwckr grkqckthph fvipyrclvg
      121 efvsdallvp dkckflhqer mdvcethlhw htvaketcse kstnlhdygm llpcgidkfr
      181 gvefvccpla eesdnvdsad aeeddsdvww ggadtdyadg sedkvvevae eeevaeveee
      241 eadddedded gdeveeeaee pyeeatertt siatttttt esveevvrvp ttaastpdav
      301 dkyletpgde nehahfqkak erleakhrer msqvmrewee aerqaknlpk adkkaviqhf
      361 qekvesleqe aanerqqlve thmarveaml ndrrrlalen yitalqavpp rprhvfnmlk
      421 kyvraeqkdr qhtlkhfehv rmvdpkkaaq irsqvmthlr viyermnqsl sllynvpava
      481 eeiqdevdel lqkeqnysdd vlanmisepr isygndalmp sltetkttve llpvngefsl
      541 ddlqpwhsfg adsvpanten evepvdarpa adrglttrpg sgltniktee isevkmdaef
      601 rhdsgyevhh qklvffaedv gsnkgaiigl mvggvviatv ivitlvmlkk kqytsihhgv
      661 vevdaavtpe erhlskmqqn gyenptykff eqmqn
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Mar 26 2007 16:06:48

Exhibit F

NCBI

Sequence Revision History

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM

Find (*Accessions*, *GI numbers* or *Fasta style SeqIds*) CAA31830 G

About Entrez

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

My NCBI (Cubby)

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Show difference in GenBank/GenPept format Show

Gi	Version	Update Date
871360	1	Nov 14 2006 11:52 PM
871360	1	Jun 23 1995 11:03 AM

```
LOCUS       CAA31830                 695 aa            linear   PRI
DEFINITION  A4 amyloid protein precursor [Homo sapiens].
ACCESSION   CAA31830
VERSION     CAA31830.1  GI:871360
DBSOURCE    embl accession X13466.1
            embl locus HSPRA42, accession X13467.1
            embl locus HSPRA43, accession X13468.1
            embl locus HSPRA44, accession X13469.1
            embl locus HSPRA45, accession X13470.1
            embl locus HSPRA46, accession X13471.1
            embl locus HSPRA47, accession X13472.1
            embl locus HSPRA48, accession X13473.1
            embl locus HSPRA49, accession X13474.1
            embl locus HSPRA410, accession X13475.1
            embl locus HSPRA411, accession X13476.1
            embl locus HSPRA412, accession X13477.1
            embl locus HSPRA413, accession X13478.1
            embl locus HSPRA414, accession X13479.1
            embl locus HSPRA415, accession X13487.1
            embl locus HSPRA416, accession X13488.1
LOCUS       871360                   695 aa            linear
DEFINITION  A4 amyloid protein precursor.
ACCESSION
VERSION       GI:871360
DBSOURCE    embl locus HSPRA41, accession X13466.6 release 44
            accession X13467
            accession X13468
            accession X13469
            accession X13470
            accession X13471
            accession X13472
            accession X13473
            accession X13474
            accession X13475
            accession X13476
            accession X13477
            accession X13478
            accession X13479
            accession X13487
            accession X13488
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Eut
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplor
            Catarrhini; Hominidae; Homo.
REFERENCE   1
```

```
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
REFERENCE   1  (residues 1 to 695)
  AUTHORS   Lemaire,H.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-NOV-1988) Lemaire H.G., Institut fuer Ger
            Universiteit zu Koeln, Weyertal 121, D-5000 Koeln 41, I
REFERENCE   2  (residues 1 to 695)
  AUTHORS   Lemaire,H.G., Salbaum,J.M., Multhaup,G., Kang,J., Bayne
            Unterbeck,A., Beyreuther,K. and Muller-Hill,B.
  TITLE     The PreA4(695) precursor protein of Alzheimer's disease
            is encoded by 16 exons
  JOURNAL   Nucleic Acids Res. 17 (2), 517-522 (1989)
   PUBMED   2783775
REFERENCE   2  (residues 1 to 695)
  AUTHORS   Lemaire,H.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-NOV-1988) Lemaire H.G., Institut fuer Ger
            Universiteit zu Koeln, Weyertal 121, D-5000 Koeln 41, I
  MEDLINE   89128427
COMMENT     On Jun 23, 1995 this sequence version replaced gi:3560(
FEATURES             Location/Qualifiers
     source          1..695
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="21"
                     /organism="unknown"
     Protein         1..695
                     /product="A4 amyloid protein precursor"
     CDS             1..695
                     /gene="PreA4"
                     /coded_by="join(X13466.1:1..57,X13467.1:452..(
                     /coded_by="join(X13466:1..57,X13467.1:452..61
                     X13468.1:12..141,X13469.1:89..201,X13470.1:11(
                     X13471.1:1..203,X13472.1:85..218,X13473.1:52..
                     X13474.1:160..318,X13475.1:1..129,X13476.1:9..
                     X13477.1:193..414,X13478.1:1..54,X13479.1:137.
                     X13487.1:179..325,X13488.1:171..272)"
                     /db_xref="GDB:119692"
                     /db_xref="GOA:P05067"
                     /db_xref="HGNC:620"
                     /db_xref="PDB:1AAP"
                     /db_xref="PDB:1AMB"
                     /db_xref="PDB:1AMC"
                     /db_xref="PDB:1AML"
                     /db_xref="PDB:1BA4"
                     /db_xref="PDB:1BA6"
                     /db_xref="PDB:1BJB"
                     /db_xref="PDB:1BJC"
                     /db_xref="PDB:1BRC"
                     /db_xref="PDB:1CA0"
                     /db_xref="PDB:1HZ3"
                     /db_xref="PDB:1IYT"
                     /db_xref="PDB:1MWP"
                     /db_xref="PDB:1OWT"
                     /db_xref="PDB:1QCM"
                     /db_xref="PDB:1QWP"
                     /db_xref="PDB:1QXC"
                     /db_xref="PDB:1QYT"
                     /db_xref="PDB:1TAW"
                     /db_xref="PDB:1TKN"
                     /db_xref="PDB:1UO7"
                     /db_xref="PDB:1UO8"
                     /db_xref="PDB:1UOA"
                     /db_xref="PDB:1UOI"
                     /db_xref="PDB:1ZE7"
                     /db_xref="PDB:1ZE9"
                     /db_xref="PDB:1ZJD"
                     /db_xref="PDB:2BEG"
                     /db_xref="PDB:2BP4"
                     /db_xref="UniProtKB/Swiss-Prot:P05067"
                     /note="pid:e"
                     /label=preA4_CDS
ORIGIN
        1 mlpglallll aawtvwalev ptdgnaglla epqiamfcgr lnmhmnvqng kv
       61 tcidtkegil qycqevypel qitnvveanq pvtiqnwckr grkqckthph fv
      121 efvsdallvp dkckflhqer mdvcethlhw htvaketcse kstnlhdygm ll
      181 gvefvccpla eesdnvdsad aeeddsdvww ggadtdyadg sedkvvevae ee
      241 eadddedded gdeveeeaee pyeeatertt siatttttt esveevvrvp tt
      301 dkyletpgde nehahfqkak erleakhrer msqvmrewee aerqaknlpk ac
      361 qekvesleqe aanerqqlve thmarveaml ndrrrlalen yitalqavpp rp
      421 kyvraeqkdr qhtlkhfehv rmvdpkkaaq irsqvmthlr viyermnqsl sl
      481 eeiqdevdel lqkeqnysdd vlanmisepr isygndalmp sltetkttve ll
```

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

Exhibit 6

JOURNAL OF
MEDICINAL
CHEMISTRY

lume 33, Number 3 March 1990

Special Topic

Molecular Modeling Software and Methods for Medicinal Chemistry[†]

N. Claude Cohen,[‡] Jeffrey M. Blaney,[§,∥] Christine Humblet,[∥] Peter Gund,[⊥] and David C. Barry[#]

Ciba-Geigy Ltd. Pharmaceutical Division, Basel, Switzerland, du Pont de Nemours & Company, Wilmington, Delaware 19898, Parke-Davis Pharmaceutical Research Division, Ann Arbor, Michigan 48105, Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065, and ICI Pharmaceuticals Division, Alderley Park, Cheshire, U.K.

Received May 11, 1989

I. Introduction

Molecular modeling has become a well-established research area during the last decade due to advances in computer hardware and software that have brought high-performance computing and graphics within the reach of most academic and industrial laboratories. A growing number of journals now focus on molecular modeling: *Journal of Computational Chemistry*, *Computers in Chemistry*, *Journal of Computer-Aided Molecular Design*, *Journal of Molecular Graphics*, *Molecular Simulations*, and *Tetrahedron Computer Methodology*. Several recent texts and reviews describe progress in molecular modeling research and applications.[1-7]

This review is intended to provide medicinal chemists with introductory material related to available molecular modeling software and methods. A particular emphasis is given to current software that integrates multiple methods, including graphic and computational tools, and focuses on systems familiar to the committee.

It is important to realize what is really meant by "computer-assisted drug design". Molecular modeling systems provide powerful tools for *building*, *visualizing*, *analyzing*, and *storing* models of complex molecular systems that can help interpret structure–activity relationships. The critical problem of *molecular design*—what structure do we build, model, and possibly synthesize?—is not answered by current methods and is left up to the creativity of the medicinal chemist. The goal of molecular modeling should not be limited only to providing insight, but it should also help to suggest new experiments, i.e., new structures tailored to have the desired biological activity. Molecular modeling cannot yet produce quantitative predictions of activity except in very special cases, but it can provide valuable qualitative guidelines that help design new lead structures. The result of a successful modeling study is therefore usually one or more candidate structures predicted to fulfill particular criteria described in a molecular model, i.e., a pharmacophore. The synthesis and biological evaluation of these target structures can be used to test and iteratively refine the model.

"Direct" and "indirect" design are the two major modeling strategies currently used in the conception of new drugs. In the first approach the three-dimensional features of a known receptor site are directly considered, and in the latter the design is based on the comparative analysis of the structural features of known active and inactive molecules that are interpreted in terms of complementarity with a hypothetical receptor site model (Figure 1). Spe-

[†] This is the second of three Special Topics on the subject of Molecular Modeling in Drug Design commissioned by the Committe on Medicinal Chemistry of IUPAC (Topliss, J. G. *J. Med. Chem.* 1988, *31*, 2229). The first article, *Guidelines for Publications in Molecular Modeling Related to Medicinal Chemistry*, appeared in an earlier issue (Gund, P.; Barry, D. C.; Blaney, J. M.; Cohen, N. C. *J. Med. Chem.* 1988, *31*, 2230). A third article on Molecular Modeling Hardware is in preparation.

[‡] Ciba-Geigy Ltd. Pharmaceutical Division.

[§] du Pont de Nemours & Company.

[∥] Parke-Davis Pharmaceutical Research Division.

[⊥] Merck Sharp & Dohme Research Laboratories.

[#] ICI Pharmaceuticals Division.

[1] Cohen, N. C. *Drugs Future* 1985, *10*, 311.

[2] Cohen, N. C. In *Advances in Drug Research*; Testa, B., Ed.; Academic Press: 1985; Vol. 14, p 41.

[3] Ripka, W. C. *Nature* 1986, *21*, 93.

[4] Burgen, A. S. V.; Roberts, G. C. K.; Tute, M. S. *Molecular Graphics and Drug Design*; Topics in Molecular Pharmacology; Vol. 3; Elsevier: Amsterdam, 1986;

[5] Gund, P.; Halgren, T. A.; Smith, G. M. *Annu. Rep. Med. Chem.* 1987, *22*, 269.

[6] Sheridan, R. P.; Venkataraghavan, R. *Acc. Chem. Res.* 1987, *20*, 322.

[7] Dean, P. M. *Molecular foundations of drug-receptor interaction*; Cambridge University Press: Cambridge, 1987.

0022-2623/90/1833-0883$02.50/0 © 1990 American Chemical Society

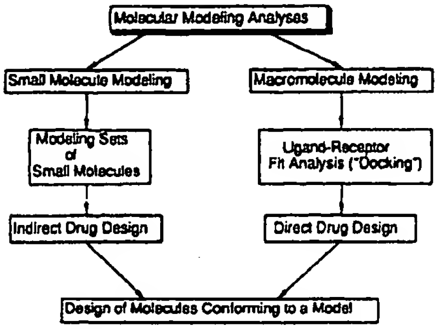
Molecular Modeling Analyses
Small Molecule Modeling
Macromolecule Modeling
Modeling Sets of Small Molecules
Ligand-Receptor Fit Analysis ("Docking")
Indirect Drug Design
Direct Drug Design
Design of Molecules Conforming to a Model

Figure 1.

cialized molecular modeling systems have been developed to analyze either the interaction of a prototype molecule with a known receptor site or the ability of a given compound to mimic the three-dimensional stereochemical features of known active compounds. Both approaches attempt to optimize receptor fit for selectivity and binding affinity while qualitatively considering other critical factors (log *P*, solubility, metabolic stability, etc.)

Most molecular modeling systems strive to provide the same basic set of features: visualization and manipulation of three-dimensional molecular models including rotatable bonds, structure building, molecular mechanics and/or dynamics, conformational analysis, electronic properties, molecular surface displays, and the calculation of various physical properties.

II. Interactive Graphics Display and Manipulation

A large range of graphics workstations are available to meet the needs of modeling applications ranging from simple, small molecule to complex macromolecules. For small molecules basic, inexpensive systems may be adequate (e.g. a Macintosh II can handle up to a couple hundred atoms in real time; real time means that the molecular model rotates and translates smoothly under interactive control). Current personal computer (PC) molecular modeling software have been reviewed recently.[175,187] The sheer size of macromolecules requires sophisticated graphics software and hardware to provide real-time, interactive response along with selective display and manipulation.[8] Current state-of-the-art systems are capable of simultaneously handling up to 20 or more molecules with up to about 20 000 atoms and thousands of molecular surface points in real time with depth-cued color and time-sliced stereo. Each molecule should be able to be individually labeled, color-coded, and controlled in three dimensions, while simultaneously monitoring inter and/or intramolecular distances and adjusting multiple contiguous or noncontinguous dihedral angles. Dials, joysticks, and a mouse, or an excellent new interactive device called "Spaceball",[9] which simultaneously control all six degrees of rotational and translational freedom with a single hand, are used to translate and rotate molecules and to rotate bonds. Typical operations are activated by first pointing to a menu and next to atoms and bonds, either with a stylus or a "mouse" to calculate, for example, distances and angles (dihedral or valence). Most systems continually update this information as the geometries are modified. The latest graphics workstations have very fast processors that do complete bump-checking (checking for contacts closer than van der Waals) and even molecular mechanics and dynamics energy calculations in real time (for small molecules up to about the size of a decapeptide). Selective control of which molecules or portions of molecules are displayed and which molecules, distances, and dihedral angles are active requires a powerful command language along with interactive "picking" of atoms and bonds with a mouse or stylus.

The trend in recent molecular modeling software design has been to exploit the powerful new windowing and computational power of the new generation of graphics workstations. This has resulted in an emphasis on menu-driven systems, which are intuitive and easy to learn, but sacrifice generality and completeness if not carefully implemented. Menu designs provide the most basic commands, but the complex syntax required by nearly all the current systems' command languages makes specifying functions not found on the menus cumbersome, if not impossible, for the nonspecialist. Hopefully, continued software design efforts will create improved menu systems and realize the need for simple, English-like command language syntax to supplement features not easily implemented in menus. The new design trend has also focused on integrating computational chemistry (e.g. molecular mechanics and dynamics) with graphics display, but much of the effort has been devoted to computations, at the expense of neglecting important features and a good user interface for interactive graphics pioneered in previous generations of graphics-only modeling systems. Despite the impressive computational performance of the new workstations, even the most sophisticated techniques provide only rough, qualitative guidance for most medicinal chemistry applications. Good interactive graphics with a well-designed user interface maximizes the performance of the most critical part of the system—the chemist.

Raster graphics has recently become the dominant technology in interactive molecular modeling, replacing the older calligraphic or vector display systems. Although raster displays have apparent advantage in providing beautiful "realistic" color solid shaded images, these images cannot be updated fast enough (with transparency and clipping) for real-time modeling yet, so vector and dot images (on raster displays) still provide the best approach for high-performance molecular modeling. Vector (bonds) and dot (molecular surface) images have the tremendous advantage of providing full transparency and clipping while displaying a complex, color-coded molecular surface and bonds, which are essential for studying interactions deep inside a macromolecular binding site or comparing several small molecules.[8] Time-sliced stereo, where the left and right eye views are alternately displayed approximately every $^1/_{30}$ s and viewed through a mechanical shutter or liquid crystal glasses synchronized to the display, provides a very convincing three-dimensional illusion and is extremely helpful for modeling complex interactions. A recent major improvement in stereo viewing is to place a liquid crystal screen over the entire graphics screen, allowing the user(s) to wear circularly polarized plastic glasses.

The simultaneous development of real-time interactive color graphics[8] and Connolly's molecular surface program[10] in 1980 revolutionized molecular modeling. Col r-coded surfaces provide qualitative displays of hydrophobic and hydrophilic regions, neutral and charged groups, electr -

[8] Langridge, R.; Ferrin, T. E.; Kuntz, I. D.; Connolly, M. L. *Science* 1981, *211*, 661.

[9] Spatial Systems Pty Ltd., PO Box 452, 65 Lavender St., Milsons Point, NSW 2061, Australia.

[10] Connolly, M. L. *Science* 1983, *221*, 709.

static potential, and mobility (based on X-ray crystallographic refinement or molecular dynamics simulation). Color-coded molecular surfaces therefore simultaneously display the main features critical to receptor binding: shape, charge, and hydrophobicity. Hydrophobic color coding was originally done simply by coloring all surface points associated with carbon "hydrophobic" (e.g. red) and all nitrogen and oxygen surface points "hydrophilic" (e.g. blue); an improved approach[11] includes "neutral" surface (e.g. yellow) for sulfur, α-carbons of amino acids, the carbon between the imidazole nitrogens in histidine, and carbonyl carbon. Molecular surfaces can also be color coded by a so-called "hydrophobic potential", based on fragment hydrophobicity values and a simple empirical function analogous to the classical formula for electrostatic potential.[12,13] Electrostatic potential molecular surfaces[14] are calculated using quantum mechanically derived partial atomic charges for each atom.[15] The potential is usually calculated one probe sphere radius above the molecular surface to give a qualitative view of what an incoming ligand "sees" as it approaches the macromolecule. The surface is color coded by the value of the electrostatic potential at each point. The electrostatic potential gradient or electric field can also be displayed graphically using short vectors.[16] Similar representations can also be envisaged for any other potential or field such as, for example, the molecular mechanics potential experienced by different chemical probes.[139]

Connolly's program[10] implemented Richard's definition[17] of molecular surface by rolling a probe sphere (usually 1.4-Å radius, the effective radius of water molecule) over the surface of the molecule, resulting in a smooth surface that represents the surface accessible to a water molecule, including internal cavities. Langridge's UCSF group[18] and Pearle and Honneger[19] independently developed van der Waals dot surface programs that are much faster than Connolly's molecular surface program, although they are not as effective at eliminating buried surface and produce a more complicated surface display for macromolecules. Both types of surface are available in most modeling systems. Connolly also developed an analytical method for calculating molecular surface,[20] which provides nearly exact values for the surface area and volume[21] enclosed by a surface along with spectacular shaded raster graphics images,[22] which gives a much different impression of a surface than the conventional CPK-like raster surfaces.[23] Barry introduced the very useful "extra radius" surface,[24] where the surface is calculated one van der Waals radius beyond the normal surface, collapsing the surface of a binding site onto the vector model of its ligand and eliminating the need for displaying the ligand's surface. This simple graphics trick makes it much easier to visualize the "docking" of a ligand into a binding site. For example, chymotrypsin's specificity for aromatic amino acid side chains is not immediately apparent from a conventional molecular surface of its active site, while the "extra radius" surface reveals an almost perfectly planar rocket that is obviously complementary to an aromatic ring. The "extra radius" surface can also be color coded by hydrophobicity or electrostatic potential.

III. Small Molecule Modeling

(a) Structure Building. Every system should provide means allowing one to construct accurate three-dimensional models of organic molecules. One of the simplest and most reliable ways is to use libraries of typical organic fragments and the Cambridge X-ray Crystallographic Data Base,[25] which contains about 50 000 structures. A molecule is constructed by assembling preexisting fragments, followed by successive adjustments of the current structure, which allows the user full control over building a reasonable starting conformation with the desired stereochemistry. Several common building functions were involved in these operations: make-bond, break-bond, fuse-rings, delete-atom, add-atom, add-hydrogens, invert chiral center, etc. They are combined with continuous refinements f the geometry of the current structure using molecular mechanics.

Most systems have facilities allowing one to draw chemical structures as a two-dimensional sketch describing the atom types (element and hybridization) and connectivity (what's bonded to what), along with some method of specifying stereochemistry (up/down, *R*/*S*, etc.). While in principle a simple and intuitive approach, it has proven very challenging to design robust methods to convert the initial two-dimensional information into reasonable low energy conformations. Most of these approaches are molecular mechanics, but often become trapped quickly in poor local minima during the conversion from two into three dimensions. Distance geometry combined with molecular mechanics[26,27] usually provides superior results to molecular mechanics alone. Very few systems are able to handle the conformational multiplicity of cyclic moieties in a fully automatic manner.[72,159] Pearlman[28] recently introduced CONCORD, an elegant method for rapidly generating good quality three-dimensional structures directly from a SMILES[29] code (a simple alphanumeric language for encoding organic structures). CONCORD is currently the best available method for generating small-molecule three-dimensional structures interactively, due to its ease of use, speed, and the quality of the resulting structure. It has the advantage of being able to produce a good quality structure for most organic compounds, including those with complex heteroatom functional groups and ring systems, without the need for developing molecular mechanics parameters. However, CONCORD

[11] Recanatini, M.; Klein, T.; Yang, C.; McClarin, J.; Langridge, R.; Hansch, C. *Mol. Pharmacol.* 1986, *29*, 436.

[12] Fauchere, J. L.; Quarendon, P.; Kaetterer, L. *J. Mol. Graphics* 1988, *6*, 203.

[13] Furet, P.; Sele, A.; Cohen, N. C. *J. Mol. Graphics* 1988, *6*, 182.

[14] Weiner, P. K.; Langridge, R.; Blaney, J. M.; Schaefer, R.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* 1982, *79*, 3754.

[15] Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* 1984, *5*, 129.

[16] Getzoff, E. D.; Tainer, J. A.; Weiner, P. K.; Kollman, P. A.; Richardson, J. S.; Richardson, D. C. *Nature* 1983, *306*, 287.

[17] Richards, F. M. *Annu. Rev. Biophys. Bioeng.* 1977, *6*, 151.

[18] Bash, P. A.; Pattabiraman, N.; Huang, C.; Ferrin, T. E.; Langridge, R. *Science* 1983, *222*, 1325.

[19] Pearl, L. H.; Honegger, A. *J. Mol. Graphics* 1983, *1*, 9.

[20] Connolly, M. L. *J. Appl. Crystallogr.* 1983, *16*, 548.

[21] Connolly, M. L. *J. Am. Chem. Soc.* 1985, *107*, 1118.

[22] Connolly, M. L. *J. Mol. Graphics* 1985, *3*, 19.

[23] Feldman, R. J.; Bing, D. H.; Furie, B. C.; Furie, B. *Proc. Natl. Acad. Sci. U.S.A.* 1978, *75*, 5409.

[24] Barry, C. D. Unpublished results.

[25] Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr.* 1979, *B35*, 2331.

[26] Wenger, J. C.; Smith, D. H. *J. Chem. Inf. Comput. Sci.* 1982, *22*, 29.

[27] Weiner, P. K.; Profeta, S., Jr.; Wipff, G.; Havel, T.; Kuntz, I. D.; Langridge, R.; Kollman, P. A. *Tetrahedron* 1983, *39*, 1113.

[28] Rusinko, A., III; Skell, J. M.; Balducci, R.; Pearlman, R. S. *CONCORD*, University of Texas at Austin; distributed by Tripos Associates, St. Louis, MO, 1987.

[29] Weininger, D. *J. Chem. Inf. Comput. Sci.* 1988, *28*, 31.

generates only a single conformer and cannot be used for conformational sampling. CONCORD has also been used to generate three-dimensional structures from two-dimensional structures stored in large industrial databases to provide conformations for newly developing three-dimensional search techniques.[30]

Many popular file formats for storing three-dimensional coordinates are in use (Brookhaven Protein Data Bank, Cambridge, Molecular Design's MOLFILE, CHEM-X' CSSR, etc.), but unfortunately there is no accepted convention or standard. The best current solution, used by more and more modeling systems to provide compatibility with other software, is to include facilities to read and write most or all of the popular formats, while making it easy for the user to add new formats. A standard molecule file format has been proposed.[160]

Molecular modeling studies result in a proliferation of files containing different results from different theoretical and experimental methods. Keeping track of all this data for several different projects can easily become a bookkeeping nightmare. Several current systems provide simple databases for storing and retrieving the results generated. A more general solution is provided by THOR,[32] an elegant chemical database system based on SMILES[29] codes. Martin et al.[33] described the use of THOR for molecular modeling.

(b) Molecular Mechanics. Molecular mechanics methods[34,35] are based on a pragmatic view of the molecular structure that is considered as a set of balls and springs with series of potential energy functions expressing the molecular force field as a sum of these functions. A typical energy equation is as follows:

$$E_{\text{total}} = E_{\text{stretching}} + E_{\text{bending}} + E_{\text{dihedral}} + E_{\text{van der Waals}} + E_{\text{electrostatic}} + E_{\text{hydrogen bond}}$$

Each of the individual energy terms have preferential equilibrium positions (bond lengths, bond angles, dihedral angles, van der Waals interaction distances, etc.) and force constants that are either experimentally known or theoretically estimated and used to associate energetic penalties with each individual deviation. A "Force Field" therefore consists of a set of analytical energy functions and their associated sets of numerical parameters. The total energy of a given molecule can be the sum of several thousands of individual contributions. Force field development remains a major problem for the large variety of complex functional groups encountered in medicinal chemistry, which is further complicated by the fact that not all force fields are readily transferable from one package to another. The most extensively tested force fields are MM2[34] (hydrocarbons plus a limited selection of simple heteroatom functional groups), AMBER[36-38] and CHARMM[39] (peptides and nucleic acids), and ECEPP[173,174] (peptid s). MM2 is the current standard for small-molecule work, but is a poor choice f r macromolecules. AMBER and CHARMM force fields are similar and are the standard for macromolecules, but give only qualitative results on small molecules. Hybrid force fields, such as the AMBER all-atom force field,[38] are usually used for calculations involving small-molecule–macromolecule interactions. Molecules that contain functional groups not parameterized by the above force fields require the estimation of new parameters specific for each new bond, bond angle, or dihedral angle type.[40] Most of the major software systems provide facilities for automatically assigning the appropriate atom types and parameters, but there is considerable variation in the quality and quantity of the parameters available. It is always prudent to calibrate unfamiliar software with some well-known test cases. Biosym[41] has formed an industrial consortium to systematically develop and test force field parameters. Assuming that all the necessary parameters are available for a given molecule, relative total strain energies can be calculated for estimating rotation or inversion barriers, preferred conformations, the energy required to achieve a specific conformation, etc. Except for special cases (e.g. estimating the enthalpy of formation of a hydrocarbon) the absolute calculated energy is of little value—*relative* energies between different conformers or isomers are important. The texts by Buckert and Allinger[34] and Clark[42] provide an excellent description of molecular mechanics and its applications.

Molecular mechanics energy minimization involves successive iterative computations, where an initial conformation is submitted to full geometry optimization. All parameters defining the geometry of the system are modified by small increments until the overall structural energy reaches a local minimum. The goal is to reach a local minimum on the potential surface within the minimum amount of time. The more sophisticated methods use the first and occasionally the second derivatives of the energy function for guiding the minimization. No method can guarantee finding the absolute lowest energy structure—the global minimum. Energy minimization will stop at the first local minimization encountered, with ut realizing that much deeper, more stable minima may be accessible. The problem is analogous to a ball rolling downhill, which stops in the first valley it finds and is unable to climb the next hill which may lead to a deeper valley. Molecular dynamics is able to climb small barriers (the barrier height depends on the temperature of the dynamics simulation) and is therefore much more efficient at locating deep local minima than simple minimizati n; short dynamics runs are now commonly used for minimization. Systematic search,[43,44] which increments all rotatable bonds in turn to explore the complete conformation space of the molecule, distance geometry[45,46] and

[30] Brint, A. T.; Willett, P. *J. Mol. Graphics* 1987, *5*, 49.

[31] Chem-X, developed and distributed by Chemical Design Ltd., Oxford, England.

[32] Weininger, D.; Weininger, A. *THOR—THesaurus ORiented chemical database, version 3.54*; Daylight Chemical Information Systems: Claremont, CA 91711, 1989.

[33] Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. *J. Comput.-Aided Mol. Des.* 1988, *2*, 15.

[34] Buckert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.

[35] Osawa, E.; Musso, H. In *Topics in Stereochemistry*; Allinger, N. L., Eliel, E. L., Wilen, S. H., Eds.; Wiley: New York, 1982; Vol. 13, p 117.

[36] Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* 1981, *2*, 287.

[37] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* 1984, *106*, 765.

[38] Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comp. Chem.* 1986, *7*, 230.

[39] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* 1983, *4*, 187.

[40] Hopfinger, A. J. *J. Comp. Chem.* 1984, *5*, 486.

[41] Biosym Technologies, Inc., 10065 Barnes Canyon Rd., San Diego, CA 92121.

[42] Clark, T. *A Handbook of Computational Chemistry*; John Wiley and Sons: New York, 1985.

[43] Dammkoehler, R. A.; Darasek, S. F.; Berkely Shands, E. F. *J. Comput.-Aided Mol. Des.* 1989, *3*, 3.

[44] Motoc, I.; Dammkoehler, R. A.; Marshall, G. R. In *Mathematical and Computational Concepts in Chemistry*; Trinajstic, N., Ed.; Horwood, Ltd.: Chichester, 1986; p 222.

other random sampling approaches attempt to locate the global minimum through thorough exploration of the allowed conformations, while the ellipsoid method[47,48] and an extension of distance geometry called energy embedding[49] can accomplish near global optimization in some cases.

Energy minimization can proceed either in internal coordinates (the variables explicitly considered are the bond lengths, bond angles, and dihedral angles) or, as is more often the case, in Cartesian coordinates (each atom is characterized with x, y, and z coordinates, and the atom moves with small increments along these axes). An advantage of minimizing in internal coordinates is that cooperative movements of several atoms or groups are well simulated in such treatments; moreover since the degrees of freedom of the chemical structures are natural, the risk that the molecules are trapped in a false minima is greatly reduced.

(c) Molecular Dynamics. In the last 10 years the static views of molecules have been considerably enlarged to include new perspectives introduced by molecular dynamics.[50,51] X-ray crystal structures represent a time-averaged structure of a continuously moving system, while molecular dynamics simulates the actual, instantaneous motion of the system. Each atom is treated as a particle responding to Newton's equations of motion: successive integrations of these equations lead to the trajectory of the atom over time in the form of a list of positions and velocities. Analyses are made through periods of typically 1–100 ps (many interesting motions are fully developed within 100 ps or less).

The motions of the atoms and chemical groups obtained by these simulations reveal subtle underlying molecular machinery and make it possible to understand phenomena that cannot be explained by the static view. Over short periods of time (e.g. a fraction of a picosecond), molecular dynamics usually shows little coherence in the displacements of the atoms. The motions are frequently interrupted by collisions with neighboring groups, and each group seems to have an erratic trajectory. Over longer periods of time, coherent and collective motions start to develop, revealing how some groups can fluctuate somewhat more than others.

The calculations require good computational power as well as appropriate graphical facilities. Animation consists of the viewing of consecutive conformations generated by molecular dynamics calculations. Animated display of molecular dynamics simulations is essential; dynamics simulations produce huge amounts of data that are difficult to interpret without graphics.

Moelcular dynamics is useful in order to identify preferred motions of either small molecules or proteins. Although it is not of direct utility in drug design except for "where does it spend most of its time" and as an improved energy minimization approach, dynamics gives a high information content picture of the precise behavior of the molecule considered and the way it can behave and interact with other partners. Restrained molecular dynamics[52] adds an artificial penalty function to restrain specific distances, angles, r dihedral angles. Restrained molecular dynamics and distance geometry[53,54] have been used to generate three-dimensional structures of small molecules, proteins, and nucleic acids consistent with NMR data.[55] Multiple energy minimization force fields are used in molecular dynamics methods and have been described in the literature.[178–186] Recent reviews[176,177] provide excellent description of molecular dynamics and related methods and illustrate various application approaches.

(d) Quantum Mechanics. In principle all treatments mentioned in the preceding paragraph can be made by using quantum chemical calculations. Molecular energies are calculated by using the Schroedinger equation with the Molecular Orbital (MO) formalism, which can provide greater accuracy along with the ability to model electronic effects not treated by molecular mechanics, as well as consume enormous amounts of computer time depending on the method and approximations used. Over a long period of time the Quantum Chemical Program Exchange (QCPE) group located at the University of Indiana has contributed greatly to the dissemination of a number of excellent theoretical chemistry programs to the scientific community.

The Schroedinger equation of a given molecular system can be solved either with no approximations at all (ab initio) or with the introduction of some approximations (semiempirical). Semiempirical treatments such as AM1,[56] MNDO,[57] CNDO[58,59] INDO,[60] EHT, MINDO,[61] PRDDO,[62] and PCILO[63,64] are some of the most popular semiempirical programs, whereas the GAUSSIAN[65] and HONDO[66] series are typical ab initio programs. AMPAC and MOPAC are QCPE packages that include the AM1, MNDO, and MINDO programs. Along with GAUSSIAN series, these are among the most popular programs for quantum mechanical calculations.[67]

(45) Crippen, G. M. *Distance Geometry and Conformational Calculations*; Bawden, D., Ed.; Research Studies Press (Wiley): New York, 1981.

(46) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Bawden, D., Ed.; Research Studies Press (Wiley): New York, 1988.

(47) Billeter, M.; Havel, T. F.; Wuthrich, K. *J. Comp. Chem.* 1987, *8*, 132.

(48) Billeter, M.; Havel, T. F.; Kuntz, I. D. *Biopolymers* 1987, *26*, 777.

(49) Crippen, G. M. *J. Phys. Chem.* 1987, *91*, 6341.

(50) Karplus, M.; McCammon, J. A. *Annu. Rev. Biochem.* 1983, *52*, 263.

(51) McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, 1987.

(52) Clore, G. M.; Nilges, M.; Brunger, A. T.; Karplus, M.; Gronenborn, A. M. *FEBS Lett.* 1987, *213*, 269.

(53) Havel, T.; Wuthrich, K. *Bull. Math. Biol.* 1984, *46*, 673.

(54) Braun, W.; Go, N. *J. Mol. Biol.* 1985, *186*, 611.

(55) Wuthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley and Sons: New York, 1986.

(56) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* 1985, *107*, 3902.

(57) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* 1977, *99*, 4899.

(58) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* 1965, *43*, S136.

(59) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* 1966, *44*, 3289.

(60) Pople, J. A.; Beveridge, D. L.; Dobosh, P. A. *J. Chem. Phys.* 1967, *47*, 2026.

(61) Bingham, R. C.; Dewar, M. J. S.; Lo, D. H. *J. Am. Chem. Soc.* 1975, *97*, 1302.

(62) Halgren, T. A.; Kleier, D. A.; Hall, J. H., Jr.; Brown, L. D.; Lipscomb, W. N. *J. Am. Chem. Soc.* 1978, *100*, 6595.

(63) Diner, S.; Malrieu, J. P.; Claverie, P. *Theor. Chim. Acta* 1969, *13*, 1.

(64) Diner, S.; Malrieu, J. P.; Jordan, F.; Gilbert, M. *Theor. Chim. Acta* 1969, *15*, 100.

(65) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; John Wiley & Sons: New York, 1986.

(66) Dupuis, M.; Rys, J.; King, H. F. *HONDO*, Quantum *Chemistry* Program Exchange, Indiana University: Bloomington, 1976.

(67) Popular programs distributed by QCPE include: MOPAC (455), AM1 (506), MNDO (428), CNDO/INDO (389), EHT (358), MINDO (309), PCILO (220), GAUSSIAN82 (446), HONDO (403), AMPAC (506).

Energies can be obtained through either the "self consistent" (SCF) formalism or with "perturbation methods". The SCF method is based on a property of the Schroedinger equation which states that whatever wave function is used to calculate the electronic energy of a given system, the corresponding energy will always be greater than the true energy value. SCF treatments are based on that property as follows: starting with an initial wave function, iteratively modify it until the total energy does not decrease. Full geometry optimizations therefore require the combination of two types of minimization: one for the calculation of the energies, and one for the optimization of the geometries.

In the perturbation methods, as in PCILO approaches,[63,64] the total energy is calculated as a convergent series of terms, with each new term improving the accuracy of the previously computed energy. The approach starts from the initial two-dimensional chemical formula that is used to compute the first term of the series. In general the treatment is stopped either at the second or at the third order. An advantage of these computations is that they are relatively rapid and permit one to obtain "conformational maps" (e.g. energy contours according to the variation of two dihedral angles). The computer time necessary to calculate a map using a 30-deg increment (12 × 12 = 144 conformations) is comparable in perturbation methods to the time necessary for only one or two conformations using SCF methods.

Quantum chemical calculations can provide detailed insight into the electronic nature of the molecular structures and allow one to analyze phenomena not yet parameterized for molecular mechanics. Molecular mechanics calculations compete favorably with MO calculations for conformational analysis and can be applied to much larger molecules; however, there are a number of physical, chemical, and electronic indices that can be obtained only with quantum mechanical treatments. These methods are theoretically powerful and can be very useful, but the tremendous amount and variety of data they generate must be interpreted with care. In some treatments, particularly when it is known that different methods might not lead to the same results, it is safer to pay more attention to the variations and the trends of the molecular property analyzed rather than to consider their absolute values. A well-known example of lack of agreement of different methods is the calculation of partial atomic charges, which are required by most molecular mechanics force fields and for the calculation of molecular electrostatic potentials. Several approaches have been developed for calculating partial atomic charges in molecules.[15,68-70] Current knowlege of the strengths and weaknesses of available semiempirical and ab initio methods was recently reviewed in an excellent introductory text.[42] Richards' text[71] provides a good introduction into applications of quantum mechanical calculations for medicinal chemistry.

In practice only molecules containing less than about 50 atoms can be studied with quantum mechanical approaches. The selection of the most appropriate method depends not only on the size of the molecule but also on the type of molecular property (e.g. conformation, electronic density, electrostatic potential, frontier orbitals, etc.) that is desired. Most major molecular modeling software packages provide interfaces to popular quantum mechanical methods.

(e) Conformational Analysis. In a first approximation, only intramolecular forces are considered to calculate the conformational properties of a given molecule. However, force field treatments are not restricted to isolated molecules ("gas phase simulations"), they can be envisaged with two molecules as in "docking" analyses, or even simulate solvent molecules in the investigation of solvent effects. Since the global energy minimum is not necessarily the receptor-bound conformation, it is essential to sample a region up to several kilocalories/mole above the global minimum. Molecular mechanics approaches are commonly used for conformational analysis, but quantum mechanical methods can be used for small molecules with two to three rotatable bonds.

A multiple conformation generation function appears now in an increasing number of modeling systems, but is often restricted to the rotation of acyclic bonds. Few modeling systems are able to handle the conformational multiplicity of cyclic (monocyclic or polycyclic) systems automatically. A robust method based on conformational assembly rules has been described[72] allowing the systematic and automatic generation of possible conformations of simple or complex cyclic molecules having, for example, precise polycyclic fused, spiro and bridge-headed systems (when the size of the rings is relatively small, e.g. less than eight members for each elementary ring). Smith et al.[73] described a variation of systematic search for cyclic systems. Gerber et al.[158] developed an elegant method for the systematic generation of conformations in macrocyclic systems that is based on generic shapes approximated by Fourier harmonic representations. More general methods based on artificial intelligence techniques were proposed to generate reliable low-energy conformations of any given small molecule.[74] Efficient variations of systematic search techniques have been described by Dammkoehler et al.[43,44] and Lipton.[75] Chang et al.[128] recently described a new Monte Carlo (random) torsion search method that appears to be one of the most efficient approaches for small molecule conformational analysis. Most major molecular modeling systems include approaches, along with extensive analysis facilities (e.g. contour plots of energy as a function of two dihedral angles). Scheraga and Colleagues have developed a series of techniques in conformational searching of polypeptides (for a review, see ref 169) that include build-up procedures,[170] increase of dimensionality,[171] Monte Carlo plus minimizations,[172] and optimization of electrostatics.[169]

Distance geometry calculations can also be used to generate random starting conformations for conformational analysis.[26,27] Distance geometry is a general method for converting a set of distance constraints into a set of three-dimensional coordinates consistent with the constraints.[45,46] The distance constraint matrix describes the complete conformation space of a molecule by including the maximum possible distance (upper bond) between each atom pair and the minimum possible distance (lower bound). All possible conformers lie between these upper and lower distance bound—distance geometry converts this distance information into three-dimensional coordinates.

[68] Pepe, G.; Serres, B.; Laporte, D.; Re, G. D.; Minichino, C. *J. Theor. Biol.* 1985, *115*, 571.

[69] Mullay, J. *J. Am. Chem. Soc.* 1986, *108*, 1770.

[70] Chirlian, L. E.; Francl, M. M. *J. Comp. Chem.* 1987, *8*, 894.

[71] Richards, W. G. *Quantum Pharmacology*, 2nd ed.; Butterworth & Co.: London, 1983.

[72] Cohen, N. C.; Colin, P.; Lemoine, G. *Tetrahedron* 1981, *37*, 1711.

[73] Smith, G. M.; Veber, D. F. *Biochem. Biophys. Res. Commun.* 1986, *134*, 907.

[74] Dolata, D. P.; Leach, A. R.; Prout, K. *J. Comput.-Aided Mol. Des.* 1987, *1*, 73.

[75] Lipton, M.; Still, W. C. *J. Comp. Chem.* 1988, *9*, 343.

Distance geometry produces a random sampling of conformation space by selecting random distances within each pair of upper and lower bounds. This approach samples conformation space rapidly and efficiently, but cannot guarantee that all of conformation space has been searched. Systematic dihedral search methods can in theory promise that all conformation space is adequately searched, but in practice, the completeness of the search is limited by the increment used in the dihedral scan. The time required for systematic search increases exponentially with each additional rotatable bond and becomes impractical beyond 12-13 rotatable bonds. The time required for distance geometry is independent of the number of rotatable bonds and depends only on the total number of atoms; distance geometry has approximately a quadratic time dependence on the number of atoms and therefore is still practical for large structures that are beyond the reach of systematic search methods. Cyclic structures are handled naturally by distance geometry with no decrease in efficiency, but systematic search method must deal with the ring-closure problem which further limits their efficiency and range.[73] Both methods require molecular mechanics calculations to calculate the energy of each generated conformation; systematic search methods often use a single-point energy calculation since bond lengths and angles are not distorted from their ideal values, but distance geometry requires at least partial energy minimization since all degrees of freedom are varied. Distance geometry is currently not available in any major molecular modeling software system, but stand-alone programs are available commercially,[76] from QCPE[53,77] or from UCSF.[78]

The ellipsoid algorithm is a promising new approach for generating low-energy conformations of molecules by efficiently sampling among the sterically allowed combinations of dihedral angles. It has been applied to the conformational analysis of 18-crown-6,[79] the determination of peptide solution structure using NMR distance constraints,[47] and ligand-protein docking.[48] For small to medium-sized molecules it may be more efficient than either systematic search or distance geometry for locating deep energy minima.

(f) Physical Properties. Although conformational analysis constitutes one important aspect of molecular modeling, a number of physical properties are also accessible with theoretical calculations. Molecular mechanics, semiempirical, and ab initio methods[42] can give rather reliable results on various molecular properties such as heats of formation, enthalpies (e.g. in evaluating the relative stability of isomers), barriers and activation energies, dipole moments, reaction paths, etc. Theoretical calculations can provide a number of indices that may not be directly related to experimental data but that can be very useful because they carry high physical information content (molecular, localized, and frontier orbitals, electronegativities, polarization, delocalization, atomic and bond population, etc.). For example, electron densities are useful because they provide a good basis for the analysis of the stereoelectronic properties of either isolated or interacting molecules. Molecular electrostatic potentials are usually generated from the partial atomic charges derived from a quantum mechanical calculation. Most of the major software systems include facilities to calculate and display electrostatic potentials. Other properties can be calculated by empirical methods; the most popular are the prediction of log *P* (octanol/water partition coefficient) and MR (molar refractivity) as developed by the Pomona College Medicinal Project.[80,81]

IV. Modeling Sets of Small Molecules

In indirect drug design the modeling is based on the recognition of three-dimensional stereochemical features common to sets of active molecules—the pharmacophore. Superposition and comparison methods, often called "molecular fitting" or "pharmacophore alignment", are the most routinely available. They compare, on a pairwise basis, an active reference compound with a set of other structures. Excluded volume analysis[82] is a classical way to geometrically compare a set of active and inactive molecules in order to reveal essential features, based on the simple idea that regions of inactive molecules which protrude beyond the volume common to the active molecules indicate sterically unfavorable regions on the receptor. The most popular approach to phamacophore superimposition has been the "active analogue" approach, developed by Marshall et al.[83,84] which uses systematic search to determine the allowed conformations of all molecules in the study, followed by comparison of interatomic distances to select conformers that overlap, based on the proposed pharmacophore. Attempts to take into consideration the conformational energies during the fitting process have been made.[85,86] The more recent "ensemble distance geometry method"[77,87] will rapidly determine if any solutions exist without replacing a complete systematic search and, if so, provide a random sampling of solutions that indicates how uniquely determined the model is. Additional advantages of this approach are that it handles rings naturally without the ring closure difficulties encountered in dihedral search methods and that chirality can be allowed to vary for any stereo centers of unknown absolute configuration.

Most available systems provide simple interactive fitting functionality by considering the molecules as conformationally rigid, while optionally allowing motion of a few dihedral angles.[85,88] Most of the major software systems

[76] Hare, D. *DSPACE*, Infinity Systems: 14810 216th Ave. NE, Woodinville, WA 98072, 1988.

[77] Blaney, J. M.; Crippen, G. M. *DGEOM*, to be submitted, Quantum Chemistry Program Exchange, Indiana University: Bloomington, 1990.

[78] Kuntz, I. D.; Crippen, G. M. *EMBED*, Department of Pharmaceutical Chemistry, University of California, San Francisco: San Francisco, CA 94143, 1980.

[79] Billeter, M.; Howard, A. E.; Kuntz, I. D.; Kollman, P. A. *J. Am. Chem. Soc.* 1988, *110*, 8385.

[80] Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley and Sons: New York, 1979.

[81] Leo, A.; Weininger, D.; Weininger, A. *CLOGP, CMR*, Medicinal Chemistry Project, Pomona College: Claremont, CA 91711; version 3.54, distributed by Daylight Chemical Information Systems, 1989.

[82] Sufrin, J. R.; Dunn, D. A.; Marshall, G. R. *Mol. Pharmacol.* 1981, *19*, 307.

[83] Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer-Assisted Drug Design*; Olson, E. C., Christofferson, R. E., Eds.; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979; p 205.

[84] Marshall, G. R.; Motoc, I. In *Molecular Graphics and Drug Design, Topics in Molecular Pharmacology*; Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Eds.; Elsevier: Amsterdam, 1986; Vol. 3, p 115.

[85] Labanowski, J.; Motoc, I.; Naylor, C. B.; Mayer, D.; Dammkoehler, R. A. *Quant. Struct.-Act. Relat.* 1986, *5*, 138.

[86] Cohen, N. C. In *Computer-Assisted Drug Design*: Olson, E. C., Christoffersen, R. E., Eds.; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979; p 371.

[87] Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. *J. Med. Chem.* 1986, *29*, 899.

[88] Cory, M.; Bentley, J. *J. Mol. Graphics* 1984, *2*, 39.

have integrated flexible fit computational modules in which not only the internal rotational degrees of freedom but also the conformational energies of the individual molecules are taken into account. MAXIMIN[88] is an example in which two alternative methods are possible: a set of flexible molecules can be mapped onto a rigid reference compound, or all the molecules are treated as flexible entities, and the treatment is directed toward the minimization of the conformational variance of the whole set. "Template forcing"[89] is another way to maximize overlaps between molecules using restrained molecular mechanics and dynamics.

In molecular fitting treatments the maximization of the overlaps is generally achieved by geometrical least-squares minimizations, which requires a preliminary selection of pairs of atoms expected to be superimposable. The choice of the pairs of atoms is very subjective, on the basis of "chamber intuition" and the hypothesized pharmacophore. Less subjective approaches have also been developed, on the basis of maximizing the overlap of a set of molecules by minimizing the exposed area of the entire set while simultaneously ensuring that the energies of the individual molecules remain close to a local minimum,[86] combinatorial methods for comparing all possible overlaps of similar atom types,[90,91] and approaches based on three-dimensional electrostatic potential similarity,[92,93] molecular surface similarity,[94] and molecular shape analyses.[161-164]

A more physical approach is to force common pharmacophore atoms to interact with a common binding site, defined by hypothetical points of interaction (e.g. dummy atoms), rather than forcing them to directly superimpose. Different chemical moieties can be compared and do not need to be exactly superimposable.[155,156] Several systems provide Boolean logical operators (and, or, not, etc.) which allow one to find common similarities between two molecules in terms of preselected electrostatic contours or molecular volumes. Cramer et al.[95] recently described a promising new 3D-QSAR method based on calculating the interaction of each molecule in a set of superimposed active structures with a variety of probe atoms on a three-dimensional lattice.

New approaches developed on databases of minimized conformers and using three-dimensional substructure and similarity search techniques[30] have already shown value in identifying pharmacophoric moieties and associated active conformations of molecules.[33] Efforts of this type are current topics of modeling development and are just now becoming available.

V. Macromolecule Modeling

X-ray crystallography and macromolecular modeling provide the most detailed possible view of drug–receptor interactions and have created a new, rational approach to drug design where the structure of a drug is designed on the basis of its fit to the three-dimensional structure in the receptor site, rather than by analogy to other active structures or random leads.[96,97] There are now over 300 X-ray crystal structures of proteins and nucleic acids that have been solved; most are available in the Brookhaven Protein Data bank,[98] including several ligand–macromolecule complexes. Although relatively few structures of actual or potential drug receptors have been solved, the rate of solving these structures has increased steadily during the last few years and will continue to increase due to improvements in crystallographic techniques and the availability of new protein through recombinant DNA approaches. Such high-resolution structures offer the potential of designing drugs tailor-made to fit their receptor with high affinity and selectivity. However, the rate of release to the public domain of three-dimensional coördinates of important macromolecules is decreasing even as the rate of solving them increases. The results of the technology that promised this great potential for rational, receptor-based drug design are in fact often not available. The issues surrounding this counterproductive situation have been discussed previously.[99,100]

Despite the impressive advances in macromolecular X-ray crystallography, availability of high-quality crystals remains the major limiting factor. 2D NMR techniques have advanced tremendously[55,101,102] and can now provide three-dimensional structural information on small proteins (up to 100–150 residues) and DNA in solution, using distance geometry[53,54] and/or restrained molecular dynamics[52,103] to build models consistent with distance constraints derived from NOE (nuclear overhauser enhancement) and coupling constant data.[55] In several cases 2D NMR has been used to solve a complete protein structure; Tendamistat, the 75-residue α-amylase inhibitor, was solved independently by 2D NMR[104,105] and X-ray crystallography,[106] resulting in very similar structures. 2D NMR previously provided only low-resolution models that revealed the overall folding pattern with little information about side-chain locations, but Wuthrich's group has recently determined the complete solution structure of Tendamistat by NMR, including all side chains[105]. The January 1989 release of the Brookhaven Protein Data Bank[98] includes for the first time a protein structure solved in solution by NMR; other structures solved by NMR will follow.

Most current software systems provide efficient means for the construction of polymeric fragments. Peptides, nucleic acids, or carbohydrates are easily generated in an arbitrary or user-defined three-dimensional conformation by selecting in a menu the linear sequence combined with additional information indicating how the progressively growing molecule should fold. The growth either can be fully extended or can follow commonly observed secondary structure (e.g. α-helix, β-sheet in the case of peptides; A,

[89] Struthers, R. S.; Rivier, J.; Hagler, A. T. In *Conformationally based drug design: Peptides and nucleic acids as templates or targets*; Vida, J. A., Gordon, M., Ed.; American Chemical Society: Washington, DC, 1984; p 239.

[90] Crippen, G. M. *J. Med. Chem.* 1980, *23*, 599.

[91] Danziger, D. J.; Dean, P. M. *J. Theor. Biol.* 1985, *116*, 215.

[92] Dean, P. M.; Chau, P. L. *J. Mol. Graphics* 1987, *5*, 152.

[93] Namasivayam, S.; Dean, P. M. *J. Mol. Graphics* 1986, *4*, 46.

[94] Dean, P. M.; Callow, P.; Chau, P. L. *J. Mol. Graphics* 1988, *6*, 28.

[95] Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* 1988, *110*, 5959.

[96] Beddell, C. R. *Chem. Soc. Rev.* 1984, *13*, 279.

[97] Hol, W. G. J. *Angew. Chem.* 1986, *25*, 767.

[98] Bernstein, F. C.; Koetzle, T. F.; Williams, G. T. B.; Meyer, E. F.; Brice, M. D.; Brodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* 1977, *112*, 535.

[99] Richards, F. M. *J. Comput.-Aided Mol. Des.* 1988, *2*, 3.

[100] Marshall, G. R.; Vinter, J. G.; Holtje, H. D. *J. Comput.-Aided. Mol. Des.* 1988, *2*, 1.

[101] Kaptein, R.; Boelens, R.; Scheek, R. M.; van Gunsteren, W. F. *Biochemistry* 1988, *27*, 5389.

[102] Wuthrich, K. *Science* 1989, *243*, 45.

[103] de Vlieg, J.; Scheek, R. M.; van Gunsteren, W. F.; Berendsen, H. J. C.; Kaptein, R.; Thomason, J. *Proteins* 1988, *3*, 209.

[104] Kline, A. D.; Braun, W.; Wuthrich, K. *J. Mol. Biol.* 1986, *189*, 377.

[105] Kline, A. D.; Braun, W.; Wuthrich, K. *J. Mol. Biol.* 1988, *204*, 675.

[106] Pflugrath, J. W.; Weigand, G.; Huber, R. *J. Mol. Biol.* 1986, *189*, 383.

B, or Z conformation for nucleic acids, and analogous prespecified conformers for carbohydrates). These simple methods have little chance of leading to meaningful three-dimensional structures unless they are used in combination with additional knowledge and experimental data.

Many more protein sequences are available than crystal structures, and the gap will continue to grow as DNA-sequencing methods become even faster. Fortunately, protein sequences occasionally show high sequence homology with proteins whose three-dimensional structure is known, suggesting the possibility of modeling the unknown structure based on the crystal structure of the homologous protein. This has become a popular approach and has recently been reviewed by Blundell et al.;[107,108] an example is the recent prediction of the three-dimensional structure of tissue plasminogen activator.[109] Homology modelling techniques have been particularly successful for predicting antibody structures.[110,111] Jones and Thirup[112] showed that it may be possible to fit most secondary structure elements using fragments from other proteins of known structure; this approach is useful for building models for insertion and deletion regions and for homology model building in general. Most of the macromolecular modeling software systems contain similar facilities for protein homology modeling.

For the majority of protein sequences with little significant homology to known structures, the problem of predicting secondary and tertiary structure accurately enough for drug design applications is still insurmountable.[113] Error rates for the various secondary structure prediction approaches are usually greater than 40%.[114,115] However, several of the current methods can suggest families of possible secondary structures that may be useful for some applications (e.g. site-directed mutagenesis). Few predictions of complete secondary and tertiary structure have been reported. A realistic appraisal of the current state of the art is represented by Cohen et al.'s ambitious prediction[116] of the core tertiary structure of Interleukin-2 prior to its determination by X-ray crystallography;[117] while the prediction had several key features correct, it was too inaccurate to be useful for drug design[118]—even small errors in the placement of secondary and tertiary structure can lead to major errors in the complete model.

VI. Modeling Drug-Receptor Interactions

The major interactions involved in drug-receptor binding are electrostatic (including hydrogen bonding), dispersion or van der Waals, and hydrophobic.[119] Hydrophobic interactions usually provide the major driving force for binding, while hydrogen-bonding and electrostatic interactions primarily provide specificity and often add little to the free energy f binding.[120-122] Drug-receptor "docking" is typically done interactively with molecular surface displays (e.g. "extra radius" surface) used to guide the fit, based on hydrophobic or electrostatic potential color coding. Since it is difficult to hit a m ving target, the binding site is usually treated as completely rigid initially, while the conformation of the ligand is adjusted interactively. Recent systems are fast enough to provide real-time energy calculations while docking (future systems may use this information to provide feedback and prevent steric collisions or high-energy conformations). High-energy contacts can be shown with color-coded vectors.[123] Interactive docking thus alternates between continuous motion, possibly with real-time updates of the interaction energy if fast hardware is available, and periodic cycles of energy minimization to clean up the visual fit. A simple feedback approach that scales the dial (or joystick) response based on the instantaneous derivative of the interaction energy facilitates docking.[124] If the user moves uphill in energy, the system resists the motion, but if the user is moving in a favorable direction, the system encourages the motion by increasing responsiveness, so the docking tends to follow the path of least resistance in a sort of interactive energy minimization. Finally, energy minimization of the entire complex, where all atoms are allowed to relax, provides a good indication of the plausibility of the model and a rough estimate of the relative interaction enthalpy of the candidate drug. Ionic interactions and hydrogen bond energies are usually overestimated in a typical calculation due to the omission of solvent hydrogen-bonding competition; these effects are treated properly in the free energy perturbation the ry method described below.

Conventional energy minimization with this many degrees of freedom is easily trapped in local minima and can give deceptive results; energy minimization rarely produces a structure that is significantly different from the starting coordinates. Molecular dynamics simulations as short as 10 ps are much better at escaping local minima and can give much lower energy structures; a good strategy is to begin with a short dynamics run and follow it with energy minimization. Such short dynamics simulations contain no meaningful information about the actual motions or dynamics of the structure (up to 30 ps may be required just for thermal equilibration); they simply provide a more efficient method of energy minimization and a good indication of the stability of the model (poor models tend to fly apart very quickly).

Multiple binding modes are often possible, as shown by the X-ray structure of an elastase-product complex in which the ligand is bound backwards to the established mode of productive binding.[125] It can be very difficult with interactive methods to find the most likely binding

[107] Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J. E.; Thornton, J. M. *Nature* 1987, *326*, 347.

[108] Blundell, T.; Carney, D.; Gardner, S.; Hayes, F.; Howlin, B.; Hubbard, T.; Overington, J.; Singh, D. A.; Sibanda, B. L.; Sutcliffe, M. *Eur. J. Biochem.* 1988, *172*, 513.

[109] Heckel, A.; Hasselbach, K. M. *J. Comput.-Aided. Mol. Des.* 1988, *2*, 7.

[110] Chothia, C.; Lesk, A. M. *J. Mol. Biol.* 1987, *196*, 901.

[111] Bruccoleri, R. E.; Haber, E.; Novotny, J. *Nature* 1988, *335*, 564.

[112] Jones, T. A.; Thirup, S. *EMBO J.* 1986, *5*, 819.

[113] Yada, R. Y.; Jackman, R. L.; Nakai, S. *Int. J. Peptide Protein Res.* 1988, *31*, 98.

[114] Kabsch, W.; Sander, C. *FEBS Lett.* 1983, *155*, 179.

[115] Nishikawa, K. *Biochim. Biophys. Acta* 1983, *748*, 285.

[116] Cohen, F. E.; Kosen, P. A.; Kuntz, I. D.; Epstein, L. B.; Ciardelli, T. L.; Smith, K. A. *Science* 1986, *234*, 349.

[117] Bandhuber, B. J.; Boone, T.; Kenney, W. C.; McKay, D. B. *Science* 1987, *238*, 1707.

[118] Landgraft, B.; Cohen, F. E.; Smith, K. A.; Gadski, R.; Ciardelli, T. L. *J. Biol. Chem.* 1989, *264*, 816.

[119] Kollman, P. A. In *X-ray Crystallography and Drug Action*; Horn, A. S., Ranter, C. J. D., Eds.; Oxford University Press: Oxford, 1984; p 63.

[120] Fersht, A. R. *TIBS* 1984, *9*, 145.

[121] Fersht, A. R.; Shi, J.; Knill-Jones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M. Y.; Winter, G. *Nature* 1985, *314*, 235.

[122] Street, I. P.; Armstrong, C. R.; Withers, S. G. *Biochemistry* 1986, *25*, 6021.

[123] Bush, B. L. *Comput. Chem.* 1984, *8*, 1.

[124] Swanson, E.; Blaney, J. M. Unpublished results.

[125] Meyer, E. F., Jr.; Radhakrishnan, R.; Cole, G. M.; Presta, L. G. *J. Mol. Biol.* 1986, *189*, 533.

mode candidates. Naruto et al.[126] used a systematic search procedure to find chymotrypsin tetrahedral intermediate conformers given a covalent bond linking the ligand with the site. DesJarlais et al.[127] developed a general docking method for conformationally flexible ligands based on a fast sphere-matching algorithm by docking each rigid fragment of the ligand (fragments between rotatable bonds) independently.

A major problem with all design approaches is our current lack of ability to calculate even a qualitatively accurate estimate of the free energy of binding between two molecules in aqueous solution. An important advance in modeling ligand–receptor interactions is the recent application of free energy perturbation methods.[129,130] This takes advantage of the properties of a thermodynamic cycle to simulate a physical process which is very difficult to calculate (the transfer of a drug from solution into a receptor binding site, compared with the transfer of its analogue) by an equivalent nonphysical process (the "mutation" of a drug into its analogue, performed both in solution and in the binding site) which is relatively easy to calculate. This "mutation" is carried out by gradually changing the parameters of the initial drug molecule to the parameters of the final drug molecule during a molecular dynamics simulation, which is performed once in "solution", usually in a box of several hundred water molecules, and again in the macromolecule. The simulation starts with 100% initial drug character and ends with 100% final drug character; intermediate steps in the simulation have nonphysical hybrid drug molecules. Molecular dynamics generates a statistical mechanical ensemble average at each point along the simulation as the properties of the initial molecule are varied. Such simulations require large amounts of supercomputer time.

Wong and McCammon[131] described the calculation of the free energy difference of binding benzamidine vs *p*-fluorobenzamide to trypsin, while Bash et al.[132] reported calculations on free energy of binding differences for several thermolysin inhibitors and for a single thermolysin inhibitor to different mutant thermolysins. Both simulations were accurate to within less than 1 kcal of the experimental value. These results demonstrated how important the role of differential solvation can be in determining binding-affinity differences. It is not clear yet how large a difference between molecules can be simulated; all drug–receptor simulations so far have involved conservative single atom replacements, although Singh et al.[133] found excellent results with changes in entire amino acid side chains for calculating differences in solvation free energy. Free-energy perturbation methods are gradually becoming available in several molecular modeling systems, although this is still a frontier research area and it is not clear what the best approaches are or how long a simulation must be run to ensure statistically significant results.

Free energy perturbation methods offer the exciting possibility of calculating accurate differences in binding free energies between related ligands, which could make it possible to predict the binding affinity of new compounds prior to synthesis. Merz and Kollman[188] recently demonstrated the predictive ability of the approach by estimating the $\Delta(\Delta G)$ of thermolysin binding to a new inhibitor. However, recent work[189,190] has pointed out that it is extremely difficult to verify when a simulation has converged and has shown that some of the early reports were rather optimistic and tended to overestimate the precision with which $\Delta(\Delta G)$ was calculated. It is now clear that additional basic research is necessary before the method can be routinely applied and yield quantitatively reliable results. Current results suggest that $\Delta(\Delta G)$ for ligand–macromolecule binding can be calculated to within ±1.5–2 kcal/mol (equivalent to about a factor of 10–30 in binding affinity). Van Gunsteren[189], and Pearlman and Kollman[190] reviewed problems and pitfalls of the approach recently.

VII. Design

In the past, drugs were designed with an almost total naivete from the point of view of the molecular mechanisms of the underlying molecular machinery involved. The recent developments in Molecular Biology have clearly revealed the critical importance of three-dimensionality (3D) in molecular recognition and discrimination aspects. Even when the 3D features of the biological proteins involved were not known, drug design conducted along with this line emerged as an important aim and stimulated the development of some of the techniques mentioned in paragraph IV. Examples of lead molecules conceived in this way have been regularly reviewed,[1,5,184] and it is beyond the scope of this article to review all the excellent contributions that were made in this perspective.

As far as direct drug design is concerned, the ability to model both small organic molecules and macromolecules in the same system is critical; several of the systems currently available were originally designed for handling the regular, repeating polymeric structure of proteins and nucleic acids and deal rather poorly with the more arbitrary structures found in small organic molecules. Others were initially designed for modeling small molecules and do not handle macromolecular structures well. Few systems come close to offering the best of macromolecular and small-molecule modeling in an integrated system, providing the ability to interactively design and build potential ligands directly into a macromolecular receptor binding site.

Computer graphics enables us to qualitatively visualize drug–receptor interactions and molecular mechanics can provide rough estimates of the interaction energy, which allow us to design molecules that are apparently complementary to a binding site. For close analogues this can be sufficient to both rationalize the relative activities of a series of analogues and design new, closely related analogues; several excellent examples of this approach have been reported.[96,97,134] An integrated approach[135] combining molecular modeling with QSAR has proven to be especially powerful for this application, since the QSAR can help differentiate between different possible binding modes. We have much less experience in the de novo design of novel molecules (without a lead compound in an X-ray structure with its receptor). The designs by Beddell et al. of 2,3-diphosphoglycerate mimics[136] and antisickling com-

[126] Naruto, S.; Motoc, I.; Marshall, G. R.; Daniels, S. B.; Sofia, M. J.; Katzenellenbogen, J. A. *J. Am. Chem. Soc.* 1985, *107*, 5262.

[127] DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. *J. Med. Chem.* 1986, *29*, 2149.

[128] Chang, G.; Still, W. C.; Guida, W. C. *J. Am. Chem. Soc.* 1989, *111*, 4379.

[129] van Gunsteren, W. F.; Berendsen, H. J. C. *J. Comput.-Aided Mol. Des.* 1987, *1*, 171.

[130] McCammon, J. A. *Science* 1987, *238*, 486.

[131] Wong, C. F.; McCammon, J. A. *J. Am. Chem. Soc.* 1986, *108*, 3830.

[132] Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* 1987, *235*, 574.

[133] Singh, U. C.; Brown, F. K.; Bash, P. A.; Kollman, P. A. *J. Am. Chem. Soc.* 1987, *109*, 1607.

[134] Roth, B. *Fed. Proc., Fed. Am. Soc. Exp. Biol.* 1986, *45*, 2765.

[135] Hansch, C.; Klein, T. *Acc. Chem. Res.* 1986, *19*, 392.

pounds[137] based on the hemoglobin X-ray structure are still some of the best examples of this approach, despite the fact that most of this work was done with wire models! The only other rep rted successful example of de novo design using computer modeling methods is the design of phospholipase A_2 inhibitors by Ripka et al.[138]

All of the approaches we have described so far are analytical and oriented toward modeling known structures. Where do the structures of novel candidate drugs come from? Actual molecular structure design is still a formidable challenge dependent on the creativity, ingenuity, and experience of the medicinal chemist. Goodford developed a simple molecular mechanics based approach for calculating optimal ligand atom locations in a binding site, which is an important first step.[139] The method is based on calculating the interaction energy for each of a variety of probes (e.g. hydroxyl oxygen, carbonyl oxygen, carboxyl oxygen, amide nitrogen, amine nitrogen, etc.) at each point on a three-dimensional grid superimposed on the binding site. The grid is then contoured by energy, and the resulting contours are graphically displayed (as color-coded contour maps or dot clouds) in the binding site. The contours indicate predicted "hot spots" where a ligand atom of a given type should prefer to bind. Unfortunately it is usually very difficult to connect each of these "hot spots" together into a synthetically accessible molecule in a low-energy conformation, but the method does provide useful visual clues for structure design.

Current design techniques combine Goodford's (or related methods) with the other previously described interactive methods, where the investigator fits a variety of organic fragments in a trial and error fashion into the site, attempting to eventually combine the fragments into a complete molecule. The best approach is usually to design and build the developing ligand piece by piece in the binding site by combining preformed fragments from a library of different ring systems and functional groups and/or with CONCORD.[23] Small molecules can be built rapidly this way, and the resulting structures are usually accurate enough for initial qualitative "docking" into the site model. This is where good interactive software design and a well-thought-out user interface are especially important, since the modeler will spend much of his time in this stage trying out new ideas. Although it seems likely that all the information required for the design of an optimal ligand is present in the high-resolution structure of the receptor site, no systematic approaches exist yet for complete de novo design. The sphere-matching flexible ligand docking approach of DesJarlais et al.[127] or a 3D pharmacophore search over a 3D database[30,33,140] may eventually be able to achieve this, by docking fragments from a large library and then combining the fragments into complete molecules.

Very recently Dean and Colleagues[165-168] have published exploratory investigations concerning the possibility of automated site-directed drug design. The aim is to conceive appropriate algorithms and to construct a knowledge base for the automatic construction of novel ligands to fit specified binding sites.

VIII. Molecular M deling S ftwar

Major currently available academic and commercial molecular modeling software systems are listed below, along with their major functions. Currently, available computer (PC) programs have limited functionality for medicinal chemistry applications; they have not been included in this paper. Gerson[175] and Sadek[187] recently compiled reviews of PC software available for basic molecular modeling applications.

Tripos[141] has developed an excellent PC (IBM PC or Apple Macintosh II) interface to the host software (running on a superminicomputer or workstation) using the PC's local processing to provide real-time graphics display and manipulation of up to 100–200 atoms. This approach, which is now appearing in an increasing number of modeling packages, takes advantage of the inexpensive, fast graphics performance of the latest generation of PC's for display of small to medium-sized molecules, but retains the full functionality of the host software on a larger computer.

Program	Functions[a]
AMBER[36]	M, MM, MD, FE
BIOGRAF[142]	G, S, M, CA, MM, MD, MO
CHEM-X[31]	G, S, M, CA, MM, STAT, MO
CONCORD[23]	S
DISGEO[53]	DG
DISMAN[54]	DG
DSPACE[76]	DG
EMBED[78]	DG
FRODO[143,144]	G, M
GRID[139,145]	PR
GROMOS[146]	M, MM, MD, FE
INSIGHT/DISCOVER/DELPHI[41]	G, S, M, CA, MM, MD, MO
MACROMODEL[147,148,157]	G, S, M, CA, MM, MD, MO
MIDAS[149,150]	G, M
MM2[34]	MM, CA
MOGLI[151]	G, S, M
QUANTA/CHARMM[39,152]	G, S, M, CA, MM, MD, FE, PR, STAT, MO
SYBYL/ALCHEMY/NITRO[141]	G, S, M, CA, MM, MD, STAT, MO

[a] *G* graphic display and manipulation
S Small molecule structure building
M Macromolecules structure building
CA Conformational analysis facilities
MM Molecular mechanics
MD Molecular dynamics
FE Free energy perturbation methods
DG Distance geometry
PR Probe interaction energies
STAT Statistical tools
MO Molecular orbital methods from QCPE

(136) Beddell, C. R.; Goodford, P. J.; Norrington, F. E.; Wilkinson, S.; Wootton, R. *Br. J. Pharmacol.* 1976, *57*, 201.

(137) Beddell, C. R.; Goodford, P. J.; Knean, G.; White, R. D.; Wilkinson, S.; Wootton, R. *Br. J. Pharmacol.* 1984, *82*, 397.

(138) Ripka, W. C.; Sipio, W. J.; Blaney, J. M. *Lect. Heterocycl. Chem.* 1987, *IX*, S95.

(139) Goodford, P. J. *J. Med. Chem.* 1985, *28*, 849.

(140) Van Drie, J. H.; Weininger, D.; Martin, Y. C. *J. Comp.-Aided. Mol. Des.*, submitted.

(141) Tripos Associates, St. Louis, MO 63117.

(142) BioDesign, 199 South Los Robles Ave., Pasadena, CA 91101.

(143) Jones, T. A. *J. Appl. Crystallogr.* 1978, *11*, 268.

(144) Jones, T. A. In *Computational Crystallography*; Sayre, D., Ed.; Clarendon Press: Oxford, 1982; p 303.

(145) Goodford, P. J. *GRID*, Molecular Discovery Ltd., West Way House, Elms Parade: Oxford OX2 9LL, England, 1986.

(146) van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS)*, Biomos: Nijenborgh 16, 9747 AG Groningen, The Netherlands, 1987.

(147) Still, W. C.; MacPherson, L. J.; Harada, T.; Callahan, J. F.; Rheingold, A. L. *Tetrahedron* 1984, *40*, 2775.

(148) Still, W. C. *Macromodel*, Department of Chemistry, Columbia University: New York, 1984.

(149) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graphics* 1988, *6*, 2.

(150) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graphics* 1988, *6*, 13.

(151) MOGLI, Evans & Sutherland Computer Corporation, Salt Lake City, UT.

IX. P rspective

Crystallographers pioneered techniques to visualize, scrutinize, and manipulate three-dimensional molecular models. For example, the ORTEP[153] program plots crystal structure illustrations. ORTEP is still widely valued, in particular to add the third-dimension perspective to molecular structure representations. Another early example of a macromolecular graphics system is FRODO,[143,144] a software program used to facilitate electron density fitting experiments and to display and examine protein structures.

Quite independently, early attempts to incorporate computational chemistry methods to study the properties of molecules of biological interest have appeared in software such as, for example, AMBER,[38] CHARMM,[39] PCILO,[63,64] MM2,[34] and CAMSEQ.[154]

It was not until later, however, that molecular modeling graphics systems emerged from the combination of the above techniques and methods. With the addition of a conformational dimension to support structure–activity studies, the medicinal chemist was progressively offered an expanding arsenal of tools to assist and enhance drug design attempts. As outlined in this review, there is now an ample choice of molecular modeling software and methods available to the medicinal chemist.

Initial modeling software packages have been designed to provide methods dedicated either to small organic molecule or macromolecular modeling applications. Recently, progress has been made in combining both applications in a single package. However, a better integration of these two aspects is still needed to improve compatibility and enhance user interaction. In addition, future developments should benefit from a concerted combination of strengths in specific techniques and methodologies, particularly when addressing the increasing number of applications for the study of the interactions between small organic molecules and macromolecules.

Recent evolution in hardware and software technologies has made possible both implementation and development of methods (e.g., molecular dynamics, real-time manipulation of colored solid-shaded images for macromolecules) that were prohibitive not so long ago. Simultaneously, software packages have progressed to take advantage of powerful state-of-the-art features (e.g., windowing, menu-driven systems, command language syntax). However, the desirable user-friendly interface has been somewhat overlooked in this evolutionary process, and modeling software can appear rather complex and cumbersome to occasional users. We hope that future developments will address this issue.

Advances in molecular modeling have been impressive over the last years. Major milestones in software and hardware technologies have been accomplished and future prospects in this rapidly evolving arena look very promising. Current efforts to develop and integrate methods and techniques to assist and enhance drug design studies should lead to even higher levels of computer automation, rationalization, quantification, and, eventually, de novo design of novel molecules.

[152] Polygen Corporation, 200 Fifth Ave., Waltham, MA 02254.

[153] Johnson, C. K. *ORTEP-II: A Fortran Thermal-Ellipsoid Plot Program for Crystal Structure Illustration*; Oak Ridge National Laboratory, ORNL-3794, UC-4-chemistry; Oak Ridge, TN, June 1965.

[154] Potenzone, R., Jr.; Cavicchi, E.; Weintraub, H. J. R.; Hopfinger, A. J. *Comput. Chem.* 1977, *1*, 187.

[155] Kato, Y.; Itai, A., Iitaka, Y. *Tetrahedron* 1987, *22*, 5229.

[156] Itai, A.; Kato, Y.; Tomioka, N.; Iitaka, Y.; Endo, Y.; Hasegawa, M.; Shudo, K.; Fujiki, H.; Sakai, S. I. *Proc. Natl. Acad. Sci. U.S.A.* 1988, *85*, 3688.

[157] Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* 1989, submitted.

[158] Gerber, P. R.; Gubernator, K.; Müller, K. *Helv. Chim. Acta* 1988, *71*, 1429.

[159] Hoflack, J.; De Clercq, P. J. *Tetrahedron* 1988, *44*, 6667.

[160] Gund, P.; Barry, D. C.; Blaney, J. M.; Cohen, N. C. *J. Med. Chem.* 1988, *31*, 2230.

[161] Hopfinger, A. J. *J. Am. Chem. Soc.* 1980, *102*, 7196.

[162] Hopfinger, A. J. *J. Med. Chem.* 1981, *24*, 818.

[163] Hopfinger, A. J. *J. Med. Chem.* 1983, *26*, 990.

[164] Hopfinger, A. J. *J. Med. Chem.* 1985, *28*, 946.

[165] Danziger, D. J.; Dean, P. M. *Proc. R. Soc. London* 1989, *236*, 101.

[166] Danziger, D. J.; Dean, P. M. *Proc. R. Soc. London* 1989, *236*, 115.

[167] Lewis, R. A.; Dean, P. M. *Proc. R. Soc. London* 1989, *236*, 125.

[168] Lewis, R. A.; Dean, P. M. *Proc. R. Soc. London* 1989, *236*, 141.

[169] Scheraga, H. A. *Prog. Clin. Biol. Res.* 1989, *289*, 3.

[170] Gibson, K. D.; Scheraga, H. A. *J. Comput. Chem.* 1987, *8*, 826.

[171] Purisima, E. O.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* 1986, *83*, 2782.

[172] Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* 1987, *84*, 6611.

[173] Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* 1975, *79*, 2361.

[174] Nemethy, G.; Pottle, M. S.; Scheraga, H. A. *J. Phys. Chem.* 1983, *87*, 1883.

[175] Gerson, C. K. *Chemical Structure Software for Personal Computers*; Meyer, D. E., Warr, W. A., Love, R. A., Eds.; American Chemical Society: Washington, DC, 1988; p 53.

[176] MacKay, D. H. J.; Cross, A. J.; Hagler, A. T. *Prediction of Protein Structure and The Principle of Protein Conformation*; Fasman, G., Ed.; Plenum Press: New York, 1989; p 317.

[177] Burt, S. K.; MacKay, D.; Hagler, A. T. *Computer-Aided Drug Design, Methods and Applications*; Perun, T. J., Propst, C. L., Eds.; Marcel Dekker, Inc.: New York, 1989; p 55.

[178] Hopfinger, A. J.; Pearlstein, R. A. *J. Comput. Chem.* 1984, *5*, 486.

[179] Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* 1986, *7*, 230.

[180] Hall, A.; Pavitt, N. *J. Comput. Chem.* 1984, *5*, 411.

[181] Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* 1985, *107*, 1489.

[182] Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* 1984, *106*, 6638.

[183] Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* 1985, *107*, 569.

[184] Boyd, D. B.; Lipkowitz, K. B. *J. Chem. Educ.* 1982, *59*, 269.

[185] Hagler, A. T.; Maple, J. R.; Thacher, T. S.; Fitzgerald, G. B.; Dinur, U. *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM Science Publishers B. V.: Leiden, 1989; p 149.

[186] Buckert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.

[187] Sadek, M.; Munro, S. *J. Comput.-Aided Mol. Des.* 1988, *2*, 81.

[188] Merz, K.; Kollman, P. A. *J. Am. Chem. Soc.* 1989, in press.

[189] van Gunsteren, W. F. *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*; van Gunsteren, W. F.; Weiner, P. K., Eds.; ESCOM Science Publishers B. V.: Leiden, 1989; p 27.

[190] Pearlman, D. A.; Kollman, P. A. ref 189, p 101.